White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of York.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/190801/

Version: Accepted Version

Article:
Venkatesh, Vilasini and Colbeck, Roger (2022) General framework for cyclic and fine-tuned causal models and their compatibility with space-time. Physical Review A. 032204. ISSN: 1094-1622

https://doi.org/10.1103/PhysRevA.106.032204

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

A general framework for cyclic and fine-tuned causal models and their compatibility with space-time

V. Vilasini[1, 2, *] and Roger Colbeck[2, †]

[1] *Institute for Theoretical Physics, ETH Zurich, 8093 Zürich, Switzerland*
[2] *Department of Mathematics, University of York, Heslington, York YO10 5DD, United Kingdom*

(Dated: 6th September, 2022)

Causal modelling is a tool for generating causal explanations of observed correlations and has led to a deeper understanding of correlations in quantum networks. Existing frameworks for quantum causality tend to focus on acyclic causal structures that are not fine-tuned i.e., where causal connections between variables necessarily create correlations between them. However, fine-tuned causal models which permit causation without correlation, play a crucial role in cryptography, and cyclic causal models can be used to model physical processes involving feedback and may also be relevant in exotic solutions of general relativity. Here we develop a causal modelling framework capable of modelling causation in these general scenarios. The key feature of our framework is that it allows operational and relativistic notions of causality to be independently defined and for connections between them to be established. The framework first gives an operational way to study causation that allows for cyclic, fine-tuned and non-classical causal influences. We then consider how a causal model can be embedded in a space-time structure (modelled as a partial order) and propose a *compatibility* condition for ensuring that the embedded causal model does not allow signalling outside the space-time future. We identify several distinct classes of causal loops that can arise in our framework, showing that compatibility with a space-time can rule out only some of them. We discuss conditions for preventing superluminal signalling within arbitrary (and possibly cyclic) causal structures and consider models of causation in post-quantum theories admitting so-called jamming correlations. Finally, this work introduces the concept of a "higher-order affects relation", which is useful for causal discovery in fined-tuned causal models.

Contents

I. Introduction 2

II. Preliminaries: Acyclic and faithful causal models 5

III. Motivation for analysing fine-tuned and cyclic causal models 7
A. Friedman's thermostat and the one-time pad 8
B. Jamming non-local correlations 9

IV. The framework, Part 1: Causality 11
A. Cyclic and fine-tuned causal models 11
B. Interventions and affects relations 13
C. Conditional and higher-order affects relations 17
D. Relationships between concepts 20

V. The framework, Part 2: Space-time 23
A. Space-time structure 23
B. Embedding of a causal model in a space-time structure 24
C. Compatibility of a causal model with an embedding in space-time 25
D. Necessary and sufficient conditions for compatibility 28

VI. Causal loops and their space-time embeddings 29
A. Different classes of causal loops 29
B. Possibility of compatibly embedding causal loops in space-time 32

*Electronic address: vilasini@phys.ethz.ch
†Electronic address: roger.colbeck@york.ac.uk

VII. Critical analysis of previous claims regarding relativistic causality 34

VIII. Summary and conclusions 36

IX. Open questions 37

Acknowledgments 40

A. Identifying conditional independences and affects relations: Examples 40
1. Jamming (Figure 9a) 41
2. Fine-tuned collider (Figure 14a) 42
3. A Type 4 affects causal loop ([1], Figure 14b) 42
4. A Type 1 affects causal loop (Figure 14c) 43

B. Further classes of affects causal loops and their space-time embeddings 43

C. Do-conditionals from causal mechanisms in quantum cyclic causal models 45

D. Proofs of all results 48
1. Proofs of Lemma IV.1 and Theorem IV.1 48
2. Proofs of Lemmas IV.3, IV.4, IV.5, IV.6, IV.7, IV.8 and Corollary IV.3 50
3. Proofs of Theorems V.1, VI.1 and Lemma VI.1 53
4. Proofs of Lemmas A.1 and A.2 54
5. Proof of Theorem B.1 55

References 57

I. INTRODUCTION

The process of identifying cause-effect relationships underlying our observations is central to science. The causal modelling paradigm [2, 3] provides mathematical tools for relating correlation and causation in scenarios described by classical variables, and have found applications in wide ranging disciplines including medical testing [4, 5], economic predictions [2, 6] and machine learning [7–9]. A consequence of Bell's theorem [10] is that in certain scenarios, classical causal models fail to explain quantum correlations [11]. This has led to a significant progress in the development of quantum causal models [11–24] that have deepened our fundamental understanding of quantum causality and quantum correlations, as well as in practical information processing tasks such as quantum cryptography, communication, quantum computation.

Previous work on quantum causality has focused on acyclic causal structures and on causal models without fine-tuned parameters, where causation and signalling become equivalent notions. While it may be considered undesirable for a physical theory of nature to allude to fine-tuned causal explanations [11], the security of cryptographic protocols such as the one-time pad rely on fine-tuning. Here, fine-tuning is required to ensure that the cipher text gives no information about the original message without the key, even though the cipher text was generated from the original message and thus causally depends on it. Cyclic causal models have been developed and widely studied in the classical causal modelling literature for describing physical scenarios with feedback [25, 26], for instance, where variables such as demand and price causally influence each other. In the quantum literature, cyclic causation has been considered in the context of more exotic phenomena such as closed timelike curves or processes with indefinite causal order [27, 28], which may be useful in approaches to quantum gravity without a definite space-time structure. The causal modelling approach enables an operational formulation of causality that is independent of space-time structure [2, 3]. Whether a cyclic causal model describes a physical scenario with feedback or a closed timelike curve depends on how the causal model is combined with space-time information (see also [29]). Thus, from a purely operational standpoint, the most general class of causal models we would like to consider include those that are cyclic, fine-tuned and also allow for non-classical causal influences. To make a connection to physical experiments, it is also desirable to characterise how this general class of causal models can be embedded in a space-time structure, such as Minkowski space-time and to characterise when they prevent violations of relativistic causality principles such as no signalling outside the future in the space-time.

In the case of acyclic causal models without fine-tuning, the condition for ensuring no superluminal signalling in a space-time is straightforward: whenever A is a cause of B in the causal model, we can interpret B as being in the future of A with respect to a space-time such as Minkowski space-time. This ensures that all causal influences and therefore

all signals propagate from past to future in the space-time. Operationally, interventions allow us to verify causation and define a notion of signalling: if intervening on A leads to different correlations on B (compared to without the intervention), then we can say that A signals to B and use this to infer that A is a cause of B. In the absence of fine-tuning, every causal relationship can be verified using interventions, and in such models, causation implies the ability to signal with an intervention. In the presence of fine-tuning, it is possible to have causation without signalling and in this case, demanding that there is no signalling outside the space-time future does not guarantee that all causal influences propagate from past to future in the space-time. The connection between superluminal signalling and causation has been previously studied by analysing correlations in Bell-type experiments in Minkowski space-time (see for instance [30, 31]). However to find conditions for ensuring no signalling outside the space-time future in arbitrary scenarios, correlations alone do not suffice; to ascertain causation we must also consider interventions. Furthermore, allowing for cyclic causal influences while considering a partially ordered space-time such as Minkowski space-time allows for an investigation of the relationships between causal loops and superluminal signalling. A mathematical framework for causally modelling these general scenarios and establishing their connection to relativistic causality principles in a space-time is currently lacking.

In this work, we develop such a framework by defining causation and space-time structure as separate notions, and then characterising their compatibility. We keep the causal part of the framework general by allowing for causation without signalling (i.e., fine-tuned causal influences), cyclic causation as well as quantum and post-quantum causes. We describe this through a causal modelling approach, but under minimal theory-independent assumptions, and while taking into account correlations as well as arbitrary interventions. We then connect this to physics by considering the embedding of the observed variables involved in the causal model into a space-time structure, such as Minkowski space, and we characterise when such embeddings do not allow superluminal signalling. The framework proposed here has two main advantages. On the one hand, keeping causation and space-time structure separate is a useful feature for considering more general formulations of physics without a fixed background space-time structure (e.g., in a theory of quantum gravity [32, 33]), while keeping a notion of processing and communicating physical information available. On the other hand, characterising the compatibility between operational causation and space-time structure can give insights into which of these scenarios is physically realisable in a space-time.

The framework introduced in this work allows a characterisation of causality in a class of post-quantum theories (producing so-called *jamming* non-local correlations) previously proposed in the literature [30, 31], clarifies the relationships between several concepts, and enables us to address a number of open questions. Even within causality conditions related to space-time, there can be several distinct notions. For example, physical principles such as "no superluminal signalling" and "no causal loops/closed time-like curves" are both associated with relativistic causality and implied by the mathematical framework of special relativity. However, these can be distinct concepts in a more general mathematical framework where the causal structure is not fully specified by the space-time structure, but only constrained by requirements such as no superluminal signalling once embedded in a space-time. Within our framework, we distinguish these concepts. In the associated Letter [1], we apply this framework to show the mathematical possibility of causal loops between Minkowski space-time events, the existence of which can be operationally detected without leading to superluminal signalling.[1] Our framework also suggests further conditions that could be used to rule out certain types of causal loops.

When we refer to operationally detectable, we mean inferences from the observed correlations and those under intervention. Some properties of an underlying causal structure can be operationally found from the observed correlations. For example, a violation of Bell inequalities within the Bell causal structure certifies the non-classicality of the underlying common cause from the observed correlations. To distinguish causation and correlation we need to consider interventions, which allow more general inferences about the causal structure [2]. Recently, it has been experimentally demonstrated [34] that the non-classicality of a causal structure can be operationally certified from causation measures based on interventions even when no such certification is possible using correlation measures alone.

Apart from these foundational implications, several features of our framework are useful from a more practical perspective. For instance, security of relativistic cryptographic protocols [35, 36] combines both relativistic notions of causality (such as the impossibility of signalling outside the future light cone) and information-theoretic concepts. Operational information about the causal structure (which encodes the structure of communication channels between agents), the embedding of the causal structure in a space-time structure, and the compatibility between the two are all relevant for cryptography.

To operationally model causation, we adopt a causal modelling approach similar to that of [2, 3], in which causal structures are represented using directed graphs. These indicate how information flows through a network of physical systems (classical, quantum or possibly those of a post-quantum probabilistic theory), and the directed graph is

[1] Here, by Minkowski space-time, we only mean the partial order corresponding to the light cone structure of Minkowski space-time.

in principle independent of any consideration of space-time. One can however consider embedding the systems represented in the causal structures within a space-time, and relativistic causality would then impose constraints on the embedding such that the causal model cannot be used to signal outside the space-time future, in which case we say that the causal model is *compatible* with the space-time structure. For example, if an active intervention on a variable A produces a change in probability distribution over another variable B, then one would say that A affects B (or A signals to B), which implies that A is a cause of B. Assigning space-time locations to the variables and requiring the effect B to always be embedded in the future light cone of the cause A makes this causal relationship compatible with the partial order of space-time. In some situations (such as for jamming [30]) we wish to allow a variable to jointly affect a set of variables without affecting individual variables in the set, and, more generally, we may consider more complicated affects/non-affects relations between arbitrary sets of variables. Such scenarios correspond to causal models where the correlations are *fine-tuned* to hide certain causal influences from direct observation such that there is causation without correlation or signalling. In the presence of fine-tuning, characterising when a causal model can be compatibly embedded in a space-time structure is more complicated. In our work we provide a method to do so by developing a general framework and introducing causal modelling tools that have applications for analysing causality in a previously proposed class of post-quantum scenarios as well more practical problems related to causal discovery, as we explain below.

Previously, minimal conditions for preventing superluminal signalling have been considered in Bell-type scenarios. This led to the introduction of a general class of post-quantum correlations that can be defined in a tripartite Bell experiment (see Figure 3) that were dubbed jamming non-local correlations [30]. In later work, the constraints defining this class of correlations were claimed to be necessary and sufficient for ruling out superluminal signalling and causal loops [31], under certain assumptions on the space-time configuration. Previous works analysing post-quantum theories admitting jamming correlations only consider the observed correlations produced in such Bell-type scenarios. However, to rigorously analyse causation and signalling possibilities in such theories, correlations alone do not suffice (since correlation does not imply causation), and interventions must also be taken into account. A defining feature of jamming correlations is that they allow the measurement setting of one party to jointly signal to the measurement outcomes of two other parties, without signalling to them individually (this can only happen with fine-tuning). In the space-time configuration considered in [30, 31], this leads to superluminal causal influences without superluminal signalling. Since we allow fine-tuning, more generally, we can consider whether it is possible to have causal loops in a causal structure that do not lead to superluminal signalling when the systems in the causal structure are embedded in Minkowski space-time. Therefore for a clear understanding of the general validity of such claims for ruling out causal loops, a rigorous causal modelling framework is required. A general framework for modelling causality and its compatibility with space-time, as described in the above paragraphs will also enable us to consider conditions for preventing signalling outside the future lightcone and causal loops in arbitrary scenarios (not just those associated with Bell experiments). To our knowledge, such a mathematical framework is lacking in the previous literature.

A framework allowing for cyclic quantum causal models was proposed in [28]. There the focus is on indefinite causal order processes and the authors adopt a fully quantum approach where all nodes correspond to quantum systems. To model post-quantum theories admitting jamming correlations [30, 31] and analyse the signalling possibilities therein, we distinguish between classical nodes corresponding to measurement settings and outcomes, and non-classical nodes (which may be quantum, or more generally post-quantum systems modelled by a generalised probabilistic theory). This is similar to the approach of [17] but, in contrast to [17], we allow for cyclic causal models, fine-tuning and also consider space-time embeddings.

Finally we note some implications for the problem of causal discovery (inferring causation from empirical data), which is ubiquitous in science. Causal discovery algorithms are often based on the assumption of "no fine-tuning" or faithfulness (see [2, 3]). Allowing fine-tuning significantly complicates causal discovery by allowing for causal influences that are not immediately reflected in certain types of empirical data. The framework, and results presented here make explicit several new aspects of fine-tuned causal models and elucidate relationships between several concepts relating to causal models that are equivalent in the absence of fine-tuning, but that become inequivalent in the presence of fine-tuning. This suggests new methods for exploring the problem of causal discovery in the presence of fine-tuning, a problem that is of interest to the scientific community beyond the foundations of quantum physics.

Summary of contributions. We first review the necessary preliminaries of the causal modelling approach in Section II and discuss the jamming scenario along with other motivating examples in Section III. In the rest of the paper, we present several results that address the open questions outlined above, which are summarised below.

- In Sections IV and V we develop an operational framework for analysing cyclic and fine-tuned causal models in the presence of latent non-classical causes, and characterising their compatibility with a space-time structure. In particular, this provides a mathematical framework for causally modelling post-quantum theories admitting jamming non-local correlations [30] (referred to as relativistic causal correlations in [31]). The framework consists

of two parts—the first concerns causal models and the second characterises the embedding of these causal models in a space-time structure.

- In the causality part of the framework (Section IV), we extend a number of results previously established in the classical causal modelling literature, typically used for acyclic and faithful causal models, to the more general scenarios considered here, such as Pearl's rules of do-calculus [2]. We also introduce several causal modelling concepts, such as "higher-order affects relations" which only become relevant in fine-tuned causal models. We derive relationships between the many distinct properties of such causal models, highlighting the deviation from the standard case of faithful causal models. These technical results have applications for the problem of causal discovery in fine-tuned causal models, which is of independent interest.

- In the second part of the framework (Section V), we use higher-order affects relations to define when a causal model can be said to be compatible with an embedding in a space-time structure, which is intended to capture that the model does not allow signalling outside the space-time future. We also consider alternative compatibility conditions (in Section V D), and discussing the relationships between them and their physical intuition.

- In Section VI, we define several distinct classes of causal loops and consider theories that are consistent with the principle that signalling outside the space-time future is not possible. We show that such theories are necessarily free of certain types of causal loops, and, in the associated letter [1], we apply our framework to construct a causal model for an operationally detectable causal loop that can be embedded in Minkowski space-time without leading to superluminal signalling. We discuss this example and illustrate in Appendix B that such theories (which allow for causal loops without signalling outside the future of a partially ordered space-time) can involve further distinct classes of causal loops beyond those defined in the main text.

- The above results illustrate the counter-intuitive possibilities allowed by fine-tuned causal models—it is logically possible to have superluminal causal influences without superluminal signalling (as in non-local hidden variable theories [37] or the jamming scenario of [30, 31]), as well as causal loops that do not lead to superluminal signalling. These results have consequences for the claim of [31] that certain conditions on correlations in a tripartite Bell scenario are necessary and sufficient for ruling out all causal loops. The claim made in [31] that certain conditions on correlations in a tripartite Bell scenario are necessary and sufficient for ruling out all causal loops does not hold in our framework without further assumptions (see Section VII).

In upcoming work [38] we apply the results of the present paper to analyse the post-quantum jamming scenario of [30, 31] in detail, where we identify an explicit protocol that leads to superluminal signalling in this setting (contrary to previous claims), as well as new properties of post-quantum theories that admit such correlations.

A reader who is more interested in the physical implications of the framework rather than causal modelling, may choose to skip the latter parts of Section IV on causal modelling, and directly move on to the space-time part of our framework in Section V. In particular, while Sections IV A and IV B are important for what follows, Examples IV.2, IV.3 and IV.4 of Section IV C already give the main intuition behind the new concept of higher-order affects relations, and how it can be applied to define compatibility with a space-time in Section V. The reader may therefore choose to skip the remaining technical details of Section IV C, as well as the subtleties of Section IV D in their first reading.

II. PRELIMINARIES: ACYCLIC AND FAITHFUL CAUSAL MODELS

We first briefly review the literature on classical and non-classical causal models, where cause-effect relationships are typically taken to be acyclic and assumed not to be fine-tuned, before developing a model where these assumptions are relaxed.

A causal structure can be represented as a directed graph over several nodes, some of which are labelled observed and some unobserved, typically this is taken to be a Directed Acyclic Graph (DAG). Each observed node corresponds to a classical random variable[2], while each unobserved node is associated with a classical, quantum or post-quantum system. A causal structure is called *classical* (denoted $\mathcal{G}^{\mathrm{C}}$), *quantum* (denoted $\mathcal{G}^{\mathrm{Q}}$) or *GPT* (denoted $\mathcal{G}^{\mathrm{GPT}}$) depending on the nature of the unobserved nodes, where GPT stands for generalised probabilistic theory [39]. Edges of causal graphs will be denoted using $\rightsquigarrow$, as it will be useful to later classify these edges into solid $\longrightarrow$ and dashed $\dashrightarrow$ ones

[2] These may represent settings or outcomes of an experiment for example.

based on certain operational conditions for detecting causation. The following definition of *cause* is implicit in the meaning of such a causal structure.

Definition II.1 (Cause)**.** Given a causal structure represented by a directed graph $\mathcal{G}$, possibly containing observed as well as unobserved nodes, we say that a node N_i is a *cause* of another node N_j if there is a directed path $N_i \rightsquigarrow \ldots \rightsquigarrow N_j$ from N_i to N_j in $\mathcal{G}$. More generally, we say that a set of nodes S_1 is a cause of a disjoint set of nodes S_2 if there exist nodes $N_i \in S_1$ and $N_j \in S_2$ such that N_i is a cause of N_j.

For an acyclic causal structure $\mathcal{G}^C$ over the n random variables $\{X_1, \ldots, X_n\}$ (i.e., having those variables as nodes), a distribution $P(X_1, \ldots, X_n)$ is said to be *compatible with* $\mathcal{G}^{\mathrm{C}}$ if it satisfies the *causal Markov condition* i.e., the joint distribution decomposes as

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | X_i^{\downarrow_1}), \tag{1}$$

where $X_i^{\downarrow_1}$ denotes the set of all parent nodes of the node X_i in the DAG $\mathcal{G}^{\mathrm{C}}$. [We later discuss a notion of compatibility for more general (possibly cyclic) causal structures (Definition IV.1), which is weaker but recovers the present definition in the classical acyclic case.] The Markov condition of Equation (1) is equivalent to the conditional independence $X_i \perp\!\!\!\perp X_i^{\nsucc} | X_i^{\downarrow_1}$ of X_i from its non-descendants, denoted $X_i^{\nsucc}$ given its parents $X_i^{\downarrow_1}$ in $\mathcal{G}$ i.e., $\forall i \in \{1, \ldots, n\}$, $P(X_i X_i^{\nsucc} | X_i^{\downarrow_1}) = P(X_i | X_i^{\downarrow_1}) P(X_i^{\nsucc} | X_i^{\downarrow_1})$ [2]. In the case of classical causal structures with unobserved nodes, the set of compatible observed distributions for the causal structure are obtained by marginalisation of a total distribution (over all nodes) that satisfies Equation (1).

In non-classical causal structures, this compatibility condition no longer applies since a node (e.g., a measurement outcome) and its parents (e.g., the quantum states that were measured to produce that outcome) in the causal structure may not coexist. Here, we can only assign a joint distribution over all the observed nodes, and this cannot in general be seen as a marginal of a joint distribution over all nodes, as in the classical case. Instead, the observed distribution in a non-classical causal structure is obtained using the states, transformations and measurements of the theory under consideration (which we will call the *causal mechanisms*), in the order specified by the causal structure and in accordance with the probability rule specified by the theory. For example, in quantum theory, this would be the Born rule. Compatibility with non-classical causal structures can be formulated in terms of a generalised Markov condition [17] that requires the non-classical causal mechanisms (e.g., the quantum channels) to factorise in a manner analogous the classical Markov condition (1), but the exact form of this will not be relevant here. There are several frameworks for describing quantum and post-quantum causal structures, which are consistent with each other and typically differ in how the nodes and edges are associated with the causal mechanisms of the theory. [For example, in the approach of [17], nodes correspond to transformations and edges correspond to propagating subsystems, while in that of [22], nodes correspond to systems and edges to channels or transformations. These details do not change the operational predictions that can be made from the causal structure, such as the possible observed correlations realisable in the causal structure and will not be needed in the rest of this paper.] As an illustration, the following example describes the sets of compatible observed correlations in the classical and quantum version of the well-known bipartite Bell causal structure $\mathcal{G}_B$ of Figure 1a. In the following, $\mathscr{P}_n$ denotes the set of all probability distributions over n random variables and $\mathscr{S}(\mathscr{H})$ denotes the set of positive semi-definite and trace one operators on a Hilbert space $\mathscr{H}$.

Example II.1 (Sets of compatible correlations in the bipartite Bell causal structure $\mathcal{G}_B$)**.** In the classical causal structure $\mathcal{G}_B^C$, the set of compatible (observed) distributions is obtained by assuming a joint distribution $P(\Lambda XYAB) \in \mathscr{P}_5$ over all nodes, that satisfies the Markov condition (1) and marginalising over the unobserved node Λ,

$$\mathscr{P}(\mathcal{G}_B^C) := \{P(XYAB) \in \mathscr{P}_4 \mid P(XYAB) = \sum_{\Lambda} P(\Lambda)P(A)P(B)P(X|A\Lambda)P(Y|B\Lambda)\}. \tag{2}$$

If Λ is a continuous random variable, the sum is replaced by an integral over Λ. This compatibility condition for the classical causal structure $\mathcal{G}_B^C$ is identical to the *local causality condition* used in the derivation of Bell inequalities (see [40] for a comprehensive review). In the quantum causal structure $\mathcal{G}_B^Q$, the unobserved node Λ corresponds to a bipartite quantum state $\rho_\Lambda \in \mathscr{S}(\mathscr{H}_\Lambda) = \mathscr{S}(\mathscr{H}_{\Lambda_X} \otimes \mathscr{H}_{\Lambda_Y})$, and the observed nodes X and Y are associated with the POVMs, $\{E_A^X\}_X$ and $\{E_B^Y\}_Y$, that act on the subsystems $\mathscr{H}_{\Lambda_X}$ and $\mathscr{H}_{\Lambda_Y}$, depending on the inputs A and B respectively to generate the output distribution.

$$\mathscr{P}(\mathcal{G}_B^Q) := \{P(XYAB) \in \mathscr{P}_4 \mid P(XYAB) = \mathrm{tr}\Big((E_A^X \otimes E_B^Y)\rho_\Lambda\Big)P(A)P(B)\}. \tag{3}$$

In classical and non-classical causal structures alike, conditional independences play an important role. For instance, in the Bell causal structure, irrespective of the nature of Λ, we have $X \perp\!\!\!\perp B|A$ and $Y \perp\!\!\!\perp A|B$. Expressed in terms of probabilities these are the no-signalling constraints. The concept of *d-separation* developed by Geiger [41] and Verma and Pearl [42] provides a method to read off implied conditional independence relations from the graph, both in classical and non-classical causal structures. It is defined as follows.

Definition II.2 (Blocked paths)**.** Let $\mathcal{G}$ be a DAG in which X and $Y \neq X$ are nodes and Z be a set of nodes not containing X or Y. A path from X to Y is said to be *blocked* by Z if it contains either $A \rightsquigarrow W \rightsquigarrow B$ with $W \in Z$, $A \leftsquigarrow W \rightsquigarrow B$ with $W \in Z$ or $A \rightsquigarrow W \leftsquigarrow B$ such that neither W nor any descendant of W belongs to Z, where A and B are arbitrary nodes in the path between X and Y.

Definition II.3 (d-separation)**.** Let $\mathcal{G}$ be a DAG in which X, Y and Z are disjoint sets of nodes. X and Y are *d-separated* by Z in $\mathcal{G}$, denoted as $(X \perp^d Y|Z)_{\mathcal{G}}$ (or simply $X \perp^d Y|Z$ if $\mathcal{G}$ is obvious from the context) if every path from a variable in X to a variable in Y is *blocked* by Z, otherwise, X is said to be *d-connected* with Y given Z.

In classical acyclic causal structures, it has been shown that every d-separation relation $X \perp^d Y|Z$ between pairwise disjoint subsets of nodes implies the corresponding conditional independence $X \perp\!\!\!\perp Y|Z$ holds for distributions compatible with the causal structure [42]. In non-classical acyclic causal structures, the same has been shown for d-separation relations between arbitrary disjoint sets of the observed nodes [17]. In our example of the Bell causal structure, we have the d-separation relations $X \perp^d B|A$ and $Y \perp^d A|B$, which imply the conditional independences $X \perp\!\!\!\perp B|A$ and $Y \perp\!\!\!\perp A|B$ characterising the no-signalling constraints.

Furthermore, in both cases, given a causal structure $\mathcal{G}$ and a distribution P compatible with it, the pair constitute a *faithful causal model* if every conditional independence $X \perp\!\!\!\perp Y|Z$ in P corresponds to a d-separation relation $X \perp^d Y|Z$ in $\mathcal{G}$. In the non-classical case, P corresponds to the distribution over the observed nodes and cannot be seen as a marginal of a joint distribution over all nodes. Hence conditional independence in the sense of $P(XY|Z) = P(X|Z)P(Y|Z)$ can only be defined when X, Y and Z are pairwise disjoint subsets of the observed nodes. In the classical case, conditional independence in this form can also be defined for unobserved nodes and in a faithful, classical causal model, all such conditional independences imply a corresponding d-separation. Note that it is possible to define a notion of conditional independence between quantum nodes in terms of conditional quantum states (instead of conditional probability distributions) [22], but in this paper, we will only consider conditional independence relations involving sets of classical variables, which could be the observed nodes of non-classical causal structures or any node of a classical causal structure. Then an *unfaithful* or *fine-tuned* causal model is one where there exists a conditional independence $X \perp\!\!\!\perp Y|Z$ in the distribution P even though X and Y are *d-connected* in $\mathcal{G}$. For example, Figure 1b provides an extension of the Bell causal structure, where there are additional causal influences from each party's input to the other party's output and it is known that any distribution realisable in the original causal structure is realisable in the classical version of this modified causal structure [11] (see Appendix C for further details). Note however that the d-separation relation $Y \perp^d A|B$ no longer holds here, and hence any no-signalling distribution would be fine-tuned or unfaithful with respect to this causal structure but not with respect to the original one of Figure 1a. In other words, the first causal structure faithfully explains no-signalling correlations using non-classical causal mechanisms while the second provides an unfaithful explanation of such correlations using classical causal mechanisms.

III. MOTIVATION FOR ANALYSING FINE-TUNED AND CYCLIC CAUSAL MODELS

One of the most common assumptions made in the analysis of causal models is that of *faithfulness* or *no fine-tuning*. Fine-tuning complicates causal inference because it involves independences that disappear with small amounts of noise, and fine-tuning is often avoided in the literature (also on the grounds that fine-tuned causal models constitute a set of measure-zero). Even in the Bell scenario explained above, a faithful explanation of the correlations using non-classical causal models is often preferred over the unfaithful explanation using classical causal models. However, there are a number of examples, as we will see below, that necessitate a fine-tuned explanation irrespective of whether the causal structure is classical and non-classical. These include certain everyday scenarios, cryptographic protocols as well as more exotic cases that arise in certain post-quantum theories that allow for superluminal influences without superluminal signalling, which we discuss in Sections III A and III B.

Another common assumption in the causality literature is that the causal structure is acyclic. Allowing fine-tuned causal influences makes possible cyclic causal structures that are compatible with minimal notions of relativistic causality, such as the impossibility of signalling superluminally at the observed level. Cyclic causal models have also found applications in the classical literature for describing systems with feedback loops [25, 43]. Developing a framework for cyclic and fine-tuned causal models in non-classical theories therefore has both foundational and practical relevance, enabling us to better understand the operational relationships between causality and signalling

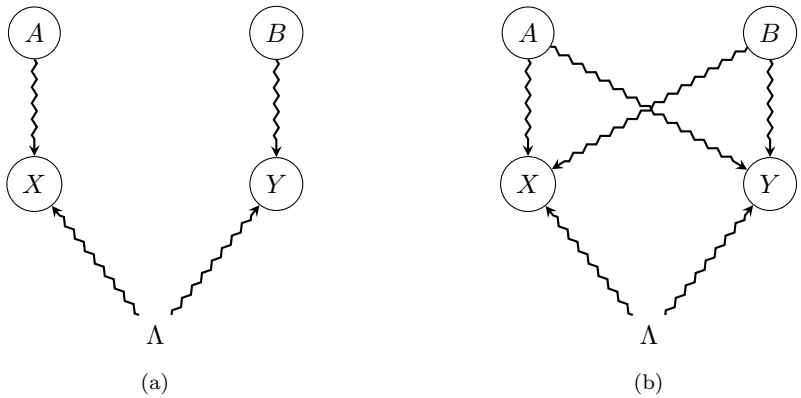


FIG. 1: (a) The bipartite Bell causal structure: Λ represents a bipartite state (classical, quantum or that of a generalised probabilistic theory) shared by two non-communicating parties Alice and Bob who measure their subsystems locally using classical measurement settings A and B to obtain classical outcomes X and Y. (b) A variation of (a) in which the settings A and B are both causes of both outcomes.

with respect to a space-time structure, and their implications for information processing. We now present some concrete examples that necessitate such causal models.

A. Friedman's thermostat and the one-time pad

Consider a house with an ideal thermostat. Such a thermostat would maintain a constant inside temperature T_I throughout the year by adjusting the energy consumption E in accordance with the outside temperature T_O. An individual who does not know how a thermostat works might conclude that T_O and E which are correlated have a causal relationship between each other while the indoor temperature T_I is causally independent of everything else. However, an engineer who is more well-versed with the workings of a thermostat knows that both T_O and E exert a causal influence on T_I, and that these influences must perfectly cancel each other out for the thermostat to function ideally. The causal model in this case is fine-tuned since the independence of T_I from T_O and E does not correspond to a d-separation relation in the causal structure (Figure 2a). This thermostat analogy which is attributed to Milton Friedman [44], can be extended to a number of other scenarios such as the effect of fiscal and monetary policies on economic growth [45], or physical systems where several forces exactly balance out.

In cryptographic settings, examples that necessitate fine-tuning include the one-time pad or the "traitorous lieutenant problem" [46]. Consider a general who wishes to relay an important secret message M to an ally and has two lieutenants available as messengers, but one of them is a traitor who might leak the message to enemies. Consider for simplicity that M is a single bit. The general could then adopt the following strategy: Depending on $M = 0$ or $M = 1$, generate two bits M_1 and M_2 such that $M_1 = M_2$ or $M_1 \neq M_2$ and with both uniformly distributed. Give M_1 to the first and M_2 to the second lieutenant to relay to the ally. Then the ally would receive M_1 and M_2 and can simply use modulo-2 addition $\oplus$ to obtain M^* which is indeed the original message $M^* = M = M_1 \oplus M_2$ (Figure 2b). More importantly, the individual messages M_1 and M_2 contain no information about M and hence neither lieutenant has any information about the secret message. A similar protocol underlies the one-time pad where a message M is encrypted using a secret key K (both binary for this example) to produce an encrypted message $M_E = M \oplus K$ which can be sent through a public channel as it will carry no information about the original message M if the key K is uniformly distributed and is kept private. Only a receiver of M_E who knows the key K can decrypt the message $M = M_E \oplus K$ (Figure 2c). Hence fine-tuning of causal influences i.e., causation in the absence of correlation, is crucial for the security.

Further, cyclic causal models have been analysed in the classical literature [25, 43] for the purpose of describing complex systems involving feedback loops, analogous to the thermostat example. Note that the cyclic dependencies here do not correspond to closed time-like curves since the variables under question are considered over a period of time— e.g., a demand at time t_1 influences the price at time $t_2 > t_1$, which in turn influences the demand at time $t_3 > t_2$. Within our framework we would use separate random variables for each of the times, which in some cases would remove the cyclicity. To characterise genuine closed time-like curves one must consider not only the pattern of causal influences, but also how the relevant variables are embedded in a space-time structure.

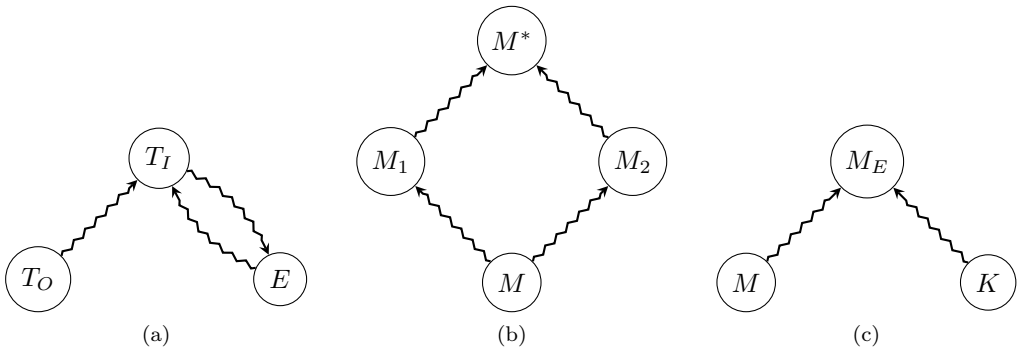


FIG. 2: **Causal structures for the motivating examples described in the main text: (a)** Friedman's thermostat **(b)** Traitorous Lieutenant **(c)** One-time pad. Note that there may be additional causal influences, for example there can be a direct influence of the outside temperature T_O and/or the inside temperature T_I on the energy consumption E in (a), the latter would make it a cyclic causal model. Further, in examples like (b), we will later see that an additional common cause between M_1 and M_2 will be required to fully explain the correlations (cf. Figure 9a).

B. Jamming non-local correlations

Another example that involves fine-tuning, even though it has not been motivated or discussed in this context, is that of jamming non-local correlations introduced in [30]. The work [30] outlines the possibility of post-quantum theories beyond the standard no-signalling probabilistic theories (such as box-world) that are still compatible with the impossibility of superluminal signalling. A better understanding of such theories would shed light on the principles of causality (beyond no superluminal signalling) that distinguish quantum and GPTs from these more general post-quantum theories. However, a mathematical framework for analysing causality in such theories is lacking, and the main purpose of this paper is to develop a general framework for modelling the relationships between causation and space-time structure, that can in particular be applied to jamming theories. In upcoming work [38], we apply our framework to the jamming scenario in more detail identifying new aspects of theories that admit such scenarios. We proceed by reviewing the jamming scenario.

Consider three space-like separated parties, Alice, Bob and Charlie sharing a tripartite system Λ which they measure using measurement settings A, B and C, producing outcomes X, Y and Z respectively. Suppose that their space-time locations are such that Bob's future light cone entirely contains the joint future of Alice and Charlie, as shown in Figure 3. The standard no-signalling conditions forbid the input of each party from being correlated with the outputs of any subset of the remaining parties, for instance, the joint distribution $P(XYZ|ABC)$ satisfies $P(XZ|ABC) = P(XZ|AC)$. In [30] it is argued that a violation of this requirement does not lead to superluminal signalling in the space-time configuration of Figure 3, as long as $P(X|ABC) = P(X|A)$ and $P(Z|ABC) = P(Z|C)$. This is because any influence that B exerts jointly (but not individually) on X and Z can only be checked when X and Z are brought together to evaluate the correlations $P(XZ|ABC)$, which is only possible in their joint future, which is by construction contained in the future of B. Bob is said to *jam* the correlations between Alice and Charlie non-locally.

In [31] the causal structure for such an experiment is represented by introducing a new random variable C_{XZ} associated with the set XZ that encodes the correlations between its elements. Then B is seen as a cause of C_{XZ} but not as a cause of either X or Z. In general scenarios, this representation would require adding a new variable for every non-empty subset of the observed nodes, which can become intractable.[3] In fact, given the assumptions that B is freely chosen and is hence a parentless node, and that for non-trivial jamming, it must be correlated with XZ, any causal structure where B is not a cause of at least one of X and Z (the causal structure proposed in [31] being such an example) would not lead to a sensible causal model satisfying the d-separation property (Definition IV.1), which is a basic property satisfied by classical and non-classical causal models alike [2, 17]. This is because such a causal structure would have a d-separation between B and XZ which would require these sets to be uncorrelated, and hence

[3] In general, this representation would include up to $2^n - 1$ observed variables whenever the original set of observed variables has n elements.

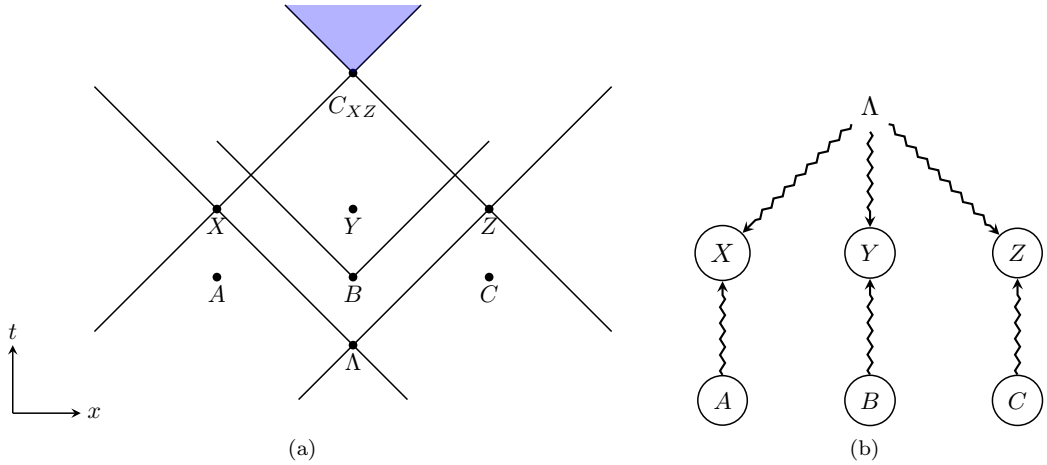


FIG. 3: **Jamming correlations in the tripartite Bell scenario:** Three parties Alice, Bob and Charlie share a tripartite system Λ, they measure their subsystem using the freely chosen measurement settings A, B and C, producing the outcomes X, Y and Z respectively, without communicating. **(a)** Space-time configuration for the jamming scenario [30, 31]: the measurement of the three parties are pairwise space-like separated with the future of Bob's input B containing the joint future of Alice's and Charlie's outputs X and Z (blue region). Here, it is argued that allowing B to signal to X and Z jointly but not individually is consistent with the principle of "no signalling outside the future lightcone", since the joint signalling can only be verified in the blue region which is in the future of B. Such correlations form a larger set as compared to the standard tripartite no-signalling correlations, which forbid individual as well as joint signalling from the inputs of any set of parties to the outputs of a complementary set of parties [49]. To model the joint signalling through jamming, a new variable C_{XZ} was introduced in [31], located at the earliest point in the joint future of X and Z and representing the correlations between X and Z. **(b)** Causal structure for the usual tripartite Bell experiment. Note that in order to explain jamming correlations in the causal modelling framework, we must either include additional causal arrows from B to X or B to Z or both (Proposition III.1), or introduce a new node C_{XZ} with an incoming arrow from B [31].

disallow any non-trivial jamming. Further, this representation does not always correspond to what is physically going on— for instance, in the example of the traitorous lieutenant, this would introduce a new variable $C_{M_1M_2}$ that is observably influenced by the general's original message M, while M would no longer be seen as a cause of M_1 or M_2. However, we know that we physically generated M_1 and M_2 using M[4], hence it is indeed a cause of at least one of them. Therefore, we aim to develop a new approach to causal modelling in a general class of fine-tuned and cyclic scenarios, using only the original variables/systems. The following proposition illustrates that the jamming scenario considered in [30, 31] necessarily corresponds to a fine-tuned causal model over the original variables. Here, jamming is considered in the context of multipartite Bell scenarios where the jamming variable is a freely chosen input of one of the parties. In the causal model approach adopted here, we will take free choice of a variable to correspond to the exogeneity of that variable in the causal structure.[5] Further, in the rest of the paper, we will denote the union $S_1 \bigcup S_2$ of any two sets S_1 and S_2 as S_1S_2.

Proposition III.1. *Consider a tripartite Bell experiment where three parties Alice, Bob and Charlie share a system Λ which they measure using the setting choices A, B and C, producing the measurement outcomes X, Y and Z respectively. Let $\mathcal{G}$ be any causal structure with only $\{A, B, C, X, Y, Z\}$ as the observed nodes where A, B and C are exogenous. Then any conditional distribution $P(XYZ|ABC)$ corresponding to the jamming correlations of [30, 31] defines a fine-tuned causal model over $\mathcal{G}$, irrespective of the nature (classical, quantum or GPT) of Λ.*

Proof. Jamming allows Bob's input B to be correlated jointly with X and Z but not individually with X or Z. Hence jamming correlations in the tripartite Bell experiment of [30, 31] are characterised by the conditions $B \perp\!\!\!\perp X$ and $B \perp\!\!\!\perp Z$ while $B \not\perp\!\!\!\perp XZ$. Since B is exogenous (i.e., has no incoming arrows), the only way to explain the correlation between

[4] And possibly some additional information to explain the distribution over the individual variables. As we will see later in Figure 9a, a common cause Λ between M_1 and M_2 would also be required in such examples.

[5] This is a standard way of modelling free choices in a causal model, although note that it is not equivalent to other definitions of free choice [31, 47, 48].

B and XZ is through an outgoing arrow or a directed path from B to the set XZ i.e., either an arrow from B to X, or from B to Z or both.[6] Since we require both independences $B \perp\!\!\!\perp X$ and $B \perp\!\!\!\perp Z$ to hold, at least one of these will not be a consequence of d-separation and hence the causal model must be fine-tuned in order to produce these correlations in the causal structure $\mathcal{G}$. □

The simplest example of a jamming is where $B = X \oplus Z$ and all variables are binary uniformly distributed (the remaining variables are irrelevant here), and we will revisit this example several times in this paper. These are the same correlations as the traitorous lieutenant example. However in the jamming case, the three variables involved are taken to be pairwise space-like separated and since B is exogenous, this corresponds to a situation where B superluminally influences the correlations between X and Z. The jamming scenario involves superluminal causal influences that need not lead to observable superluminal signalling. Generalising from this idea, one can consider whether such influences can be used to create causal loops that do not lead to any signalling to the past, or even outside the space-time future. In the interest of generality and understanding the relationships between the principles of "no superluminal signalling" and "no causal loops", one must consider fine-tuned causal influences along with cyclic causal influences, and characterise when these influences may or may not lead to signalling outside the future with respect to a space-time structure, even in the presence of latent non-classical causes.

IV. THE FRAMEWORK, PART 1: CAUSALITY

This section is devoted to outlining our causal modelling framework. Section IV A provides a minimal definition (Definition IV.1) of a causal model, allowing cyclic, fine-tuned and non-classical causal influences, including when an observed distribution is compatible with a causal structure. In Section IV B, we describe the use of interventions within such causal models. This enables us to show that Pearl's rules of do-calculus [2] hold in the more general causal models defined here (Theorem IV.1). Interventions give rise to *affects relations* which capture the notion of signalling in a causal model (Definition IV.3). Using these we classify the causal arrows in terms of whether or not they enable signalling. For some of our results we find it useful to extend these affects relations to *conditional and higher-order (HO) affects relations* (Section IV C), which capture the most general way of signalling in our framework, through joint interventions on multiple nodes. Corollary IV.3 gives a main implication of conditional HO affects relations on the underlying causal structure. Section IV D summarises the relations between the various concepts and illustrates them with several examples.

A. Cyclic and fine-tuned causal models

Following the motivation set out in the previous sections, we wish to relax the assumptions of acyclicity and faithfulness and extend causal modelling methods to cyclic and fine-tuned causal structures with latent quantum and post-quantum causes. While quantum cyclic causal models have been previously studied [28], these have been analysed in the faithful case and are based on the split-node causal modelling approach of [22]. This approach is not equivalent to the standard causal modelling approach such as [17] in the cyclic case, for example the former forbids faithful 2 node cyclic causal structures [28] but the latter does not, and the former admits a Markov factorisation (analogous to Equation (1)) while the latter does not in general (as explained in the next paragraph). To the best of our knowledge, there is no prior framework for causally modelling cyclic and unfaithful causal structures in the presence of quantum and post-quantum latent nodes, the lack of a Markov factorisation posing a particular challenge. Here, we propose a framework for achieving this.[7] We will define causal models in terms of minimal conditions that they must satisfy at the level of the observed nodes which are classical.

a. Observed distribution: In classical acyclic causal models, the causal Markov condition (1) is used for defining the compatibility of the observed distribution with the causal structure [2]. In the non-classical case, an analogous generalised Markov condition of [17] constraining the non-classical causal mechanisms (states, transformations and measurements) provides a compatibility condition. However, in cyclic causal models, demanding such a factorisation will be too restrictive even in the classical case. For example, consider the simplest cyclic causal structure, the 2-cycle where $X \rightsquigarrow Y$ and $Y \rightsquigarrow X$, with X and Y observed and $X = Y$. Used naïvely, the Markov condition would imply

[6] If this were not the case, B would be d-separated from XZ and therefore cannot be correlated with it.

[7] Note that there may be other, inequivalent ways to do the same, based on a different condition for compatibility of a distribution with a causal structure, for example.

that $P(XY) = P(X|Y)P(Y|X)$. Since $X = Y$, the right hand side is a product of deterministic distributions, which forces $P(XY)$ to also be deterministic in order to be a valid distribution. In order to not restrict directed cycles to only consist of deterministic variables, we instead use a weaker compatibility condition in terms of d-separation between observed nodes. As we have previously noted, this is a concept that also applies to non-classical causal structures. The condition captures the intuition that certain graph separation properties in the causal structure must imply (conditional) independences in the correlations it gives rise to. Based on this, we define compatibility of the observed distribution with a cyclic causal structure as follows within our framework.

Definition IV.1 (Compatibility of observed distribution with a causal structure)**.** Let $\{X_1, \ldots, X_n\}$ be a set of random variables denoting the observed nodes of a directed graph $\mathcal{G}$ (which may also have unobserved nodes), and $P(X_1, \ldots, X_n)$ be a joint probability distribution over them. Then P is said to be *compatible with* $\mathcal{G}$ (or to *satisfy the d-separation property*) if for all disjoint subsets X, Y and Z of $\{X_1, \ldots, X_n\}$,[8]

$$X \perp^d Y|Z \quad \Rightarrow \quad X \perp\!\!\!\perp Y|Z \quad \text{i.e., } P(XY|Z) = P(X|Z)P(Y|Z).$$

In the previous literature, causal models are typically defined in terms of a causal structure and causal mechanisms (which are then used to derive the observed distribution). When doing so it known that Definition IV.1 is satisfied by classical as well as non-classical causal models in the acyclic case [2, 17]. The compatibility property holds in several classical cyclic causal models [25, 43]. For classical acyclic models, it is equivalent to the causal Markov condition (1) [51]. In Appendix C, we provide an example of a quantum cyclic causal model (with causal mechanisms) where this holds. However there also exist cyclic causal models producing observed distributions that do not satisfy Definition IV.1, we discuss this further in the Appendix as well. There, we also present further motivation for the compatibility condition of Definition IV.1 in terms of the properties of the underlying causal mechanisms (e.g., functional dependences in the classical case or completely positive maps in the quantum case) and outline possible methods for identifying when this condition might hold for non-classical cyclic causal models. Even in the classical case, several inequivalent definitions of compatibility are possible (which become equivalent in the acyclic case) and [25] presents a detailed analysis of these conditions and the relationships between them. Such an analysis for the non-classical case is beyond the scope of the present work. For the rest of this paper, we will only consider causal models that satisfy the compatibility condition IV.1.

We will work with the following minimal definition of a causal model in this paper which is in terms of the graph and observed distribution only. Further details about the causal mechanisms such as the functional relationships between classical variables, choice of quantum states or transformations, or generalised tests [17] can also be included in the full specification of the causal model. These constitute the *causal mechanisms* of the model. Developing a complete and formal specification of these mechanisms and deriving the conditions for their compatibility with cyclic, fine-tuned and non-classical causal models is a tricky problem, we outline possible ideas for this in Appendix C and leave the full problem for future work. The results of this paper hold without such a specification which if added would be a way to generalise them. Interestingly, we find that even with this minimal definition, we can derive several new results for a general class of causal models and also reproduce results that were originally derived for acyclic classical causal models.

Definition IV.2 (Causal model)**.** A causal model over a set of observed random variables $\{X_1, \ldots, X_n\}$ consists of a directed graph $\mathcal{G}$ over them (possibly involving classical, quantum or GPT unobserved systems) and a joint distribution $P_{\mathcal{G}}(X_1, \ldots, X_n)$ that is compatible with the graph $\mathcal{G}$ according to Definition IV.1.

Note that other definitions of causal model are used in the literature, in particular, sometimes the definition requires that $P_{X_i|\text{par}(X_i)}$ (or more generally, a possibly non-classical channel from $\text{par}(X_i)$ to X_i) is given for each node X_i, see e.g. [2, 17].

Definition IV.1 allows for fine-tuned distributions to be compatible with the causal structure since it only requires that d-separation implies conditional independence and not the converse. Fine-tuned causal models may in general have an arbitrary number of additional conditional independences that are not implied by the d-separation relations in the corresponding causal graph. The following lemma shows that some additional conditional independences that are not directly implied by d-separation can be derived using d-separation and other independences (not implied by d-separation) that may be provided.

[8] Note that we only need to consider d-separation between observed sets of variables in this definition, however the paths being considered may involve unobserved nodes. For example, if the observed variables X and Y have an unobserved common cause Λ, then X and Y are not d-separated by the empty set since there is an unblocked path between X and Y through the unobserved common cause, and naturally we do not expect X and Y to be independent in this case.

Lemma IV.1. *Let S_1, S_2 and S_3 be three disjoint sets of RVs such that $S_1 \perp\!\!\!\perp S_2|S_3$. If S is a set of RVs that is d-separated from these sets in a directed graph $\mathcal{G}$ containing all the members of S_1, S_2, S_3 and S as nodes i.e., $S \perp^d S_i$ $\forall i \in \{1,2,3\}$, then any distribution P that is compatible with $\mathcal{G}$ also satisfies the following conditional independences,*

$$S_1 S \perp\!\!\!\perp S_2|S_3, \quad S_1 \perp\!\!\!\perp S_2 S|S_3 \text{ and } \quad S_1 \perp\!\!\!\perp S_2|S_3 S.$$

A proof can be found in Appendix D 1. Note that this lemma is trivial in the case of faithful causal models. This is because, the independence $S_1 \perp\!\!\!\perp S_2|S_3$ implies the d-separation $S_1 \perp^d S_2|S_3$ for a faithful causal model. Then, combined with $S \perp^d S_i$, we get the d-separations $S_1 S \perp^d S_2|S_3$, $S_1 \perp^d S_2 S|S_3$ and $S_1 \perp^d S_2|S_3 S$, which in turn imply the corresponding independences. This property is not so straighforward for fine-tuned causal models but nevertheless holds. Specific examples of this property for fine-tuned causal models are discussed in Appendix A.

B. Interventions and affects relations

So far, we have only discussed the possible correlations that can be compatible with a causal structure. However, it is not possible to infer an underlying causal structure from correlations alone: correlations are symmetric while causal relationships are directional. For example, if two variables X and Y are correlated, Reichenbach's principle [50] asserts that either X must be a cause of Y, Y must be a cause of X, X and Y share a common cause or any combination thereof. These causal explanations cannot be distinguished on the basis of observed correlations alone. However, intuitively, we can argue that if "doing" something only to X produces a change in the distribution over Y, then X is a cause of Y. This intuition is formalised in terms of interventions and do-conditionals [2], and we will adopt the augmented graph approach [2] for defining these.

a. Pre-intervention, augmented and post-intervention causal strutures: Consider a causal model associated with a causal structure $\mathcal{G}$ over a set $S = \{X_1, \ldots, X_n\}$ of observed nodes. External intervention on a node $X \in S$ can be described using an augmented graph $\mathcal{G}_{I_X}$ which is obtained from the original graph $\mathcal{G}$ by adding a node I_X and an edge $I_X \rightsquigarrow X$ (with everything else unchanged). The intervention variable I_X can take values in the set $\{\text{idle}, \{\text{do}(x)\}_{x \in X}\}$, where $I_X = \text{idle}$ corresponds to the case where no intervention is performed (i.e., the situation described by the original causal model) and $I_X = \text{do}(x)$ forces X to take the value x by cutting off its dependence on all other parents. From this, we see that whenever $I_X \neq \text{idle}$, X no longer depends on its original parents $\text{par}_{\mathcal{G}}(X)$. Therefore, conditioned on $I_X \neq \text{idle}$, it is illustrative to consider a new graph which we denote by $\mathcal{G}_{\text{do}(X)}$ that represents the post-intervention causal structure after a non-trivial intervention has been performed. The causal graph $\mathcal{G}_{\text{do}(X)}$ is obtained by cutting off all incoming arrows to X except the one from I_X in the causal graph $\mathcal{G}_{I_X}$, with everything else unchanged. An example of the graphs $\mathcal{G}$, $\mathcal{G}_{I_X}$ and $\mathcal{G}_{\text{do}(X)}$ is given in Figure 4. The above procedure also applies to interventions on subsets of the nodes, for example, if X is a subset of the observed nodes that is being intervened on, an exogenous intervention variable I_{X_i} will be introduced for each element X_i of X, along with the corresponding edge $I_{X_i} \rightsquigarrow X_i$. Then, $I_X \rightsquigarrow X$ will be used as a short hand to denote that each element of $I_X = \{I_{X_i}\}_i$ has a direct causal arrow to the corresponding X_i. Note that requiring each I_{X_i} to be exogenous ensures that the intervention to be performed on each node is chosen independently (in principle, one could consider correlated interventions as well but we do not do so here).

b. Defining the post intervention causal model: The effect of an intervention on the node X setting $X = x$, i.e., performing $\text{do}(x)$ is to transform the original probability distribution $P_{\mathcal{G}}(X_1, \ldots, X_n)$ into a new probability distribution $P_{\mathcal{G}_{\text{do}(X)}}(X_1, \ldots, X_n, I_X)$. These distributions are compatible with the original (i.e., pre-intervention) and the post-intervention graphs, $\mathcal{G}$ and $\mathcal{G}_{\text{do}(X)}$ respectively and the following defining rules tell us some of the relationships between these distributions. Here the distribution $P_{\mathcal{G}_{I_X}}(X_1, \ldots, X_n, I_X)$ compatible with the augmented graph $\mathcal{G}_{I_X}$ mediates the relationships between the pre and post intervention scenarios. Note that the set of intervention variables I_X is additionally introduced in going from $\mathcal{G}$ to $\mathcal{G}_{I_X}$ or $\mathcal{G}_{\text{do}(X)}$. In the corresponding causal models, the distribution over I_X can be arbitrary and all of the following definitions and results hold for any choice of P_{I_X}. Then, for any two disjoint subsets X and Y of the observed nodes, the following defining equations hold.

$$P_{\mathcal{G}_{I_X}}(Y|I_X = \text{idle}) = P_{\mathcal{G}}(Y) \tag{4a}$$

$$P_{\mathcal{G}_{I_X}}(Y|I_X = \text{do}(x)) = P_{\mathcal{G}_{\text{do}(X)}}(Y|I_X = \text{do}(x)) = P_{\mathcal{G}_{\text{do}(X)}}(Y|X = x) \quad \forall x \tag{4b}$$

$$P_{\mathcal{G}_{I_X}}(Y|I_X = \text{do}(x), X = x) = P_{\mathcal{G}_{I_X}}(Y|I_X = \text{do}(x)) \quad \forall x \tag{4c}$$

$$P_{\mathcal{G}_{I_X}}(I_X = \text{do}(x), X = x') = 0 \quad \forall x, x' \text{ such that } x \neq x' \tag{4d}$$

Intuitively, the first equation tells us that when all the intervention variables are "idle", this corresponds to the original causal model, as no intervention is performed. The remaining three equations capture the fact that when a

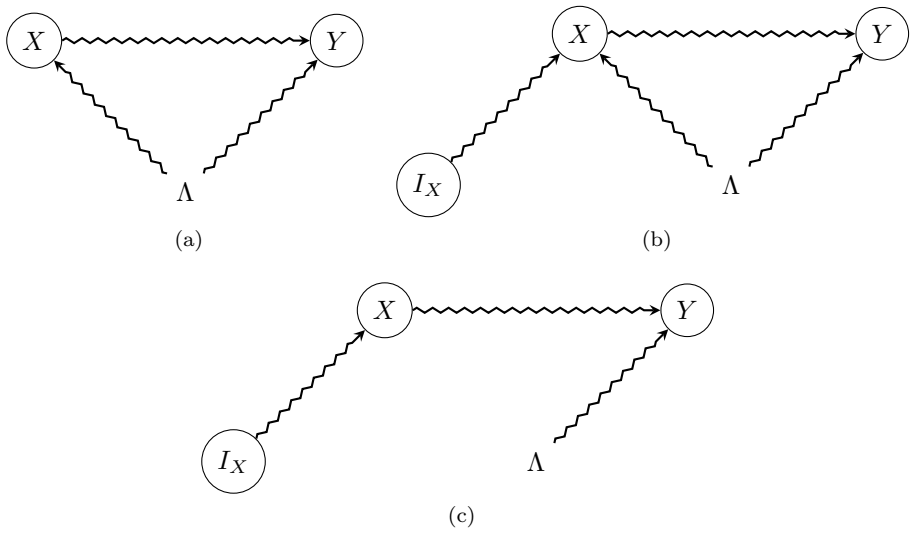


FIG. 4: **Pre-intervention, augmented and post-intervention causal structures:** Taking the original, pre-intervention causal structure, $\mathcal{G}$, to be that of (a), parts (b) and (c) of this figure illustrate the augmented causal structure, $\mathcal{G}_{I_X}$, and post-intervention causal structure, $\mathcal{G}_{\mathrm{do}(X)}$, for intervention on X. In $\mathcal{G}_{I_X}$, the variable I_X can take values in the set $\{\mathrm{idle}, \{\mathrm{do}(x)\}_{x\in X}\}$ while in $\mathcal{G}_{\mathrm{do}(X)}$, it can only take the values $\{\mathrm{do}(x)\}_{x\in X}$ corresponding to an active intervention. Conditioned on $I_X = \mathrm{idle}$, we effectively obtain the original causal model (a) which corresponds to no intervention being performed, as specified by Equation (4a).

non-trivial intervention is performed, each intervention variable $I_{X_i} \in I_X$ is perfectly correlated with the corresponding intervened variable $X_i \in X$. The conditional probability distribution $P_{\mathcal{G}_{\mathrm{do}(X)}}(Y|X = x)$ of Equation (4b) is often denoted simply as $P(Y|\mathrm{do}(x))$ and commonly referred to as the *do-conditional*. Note that $P(y|\mathrm{do}(x)) := P_{\mathcal{G}_{\mathrm{do}(X)}}(y|x) \neq P(y|x) := P_{\mathcal{G}}(y|x)$ in general. At first sight, it might appear that these defining equations do not tell us how the pre and post intervention distributions $P_{\mathcal{G}}$ and $P_{\mathcal{G}_{\mathrm{do}(X)}}$ are related since $P_{\mathcal{G}}$ is related to $P_{\mathcal{G}_{I_X}}$ only when $I_X = \mathrm{idle}$ (Equation (4a)) and $P_{\mathcal{G}_{I_X}}$ is related to $P_{\mathcal{G}_{\mathrm{do}(X)}}$ only when $I_X \neq \mathrm{idle}$. However, as we will see in subsequent sections, these defining rules along with compatibility condition of Definition IV.1 allow us to derive further useful rules that explicitly connect the pre and post intervention distributions. The intuition for this is that the augmented and post-intervention graphs are constructed from the pre-intervention graph and certain d-separations in the pre-intervention graph imply corresponding d-separations in the augmented and post-intervention graphs, and therefore certain independences in the associated distributions.

c. The physical picture: At the level of the causal mechanisms (if these are also given), the causal mechanisms of $\mathcal{G}_{\mathrm{do}(X)}$ can be obtained from those of $\mathcal{G}$ simply by updating the causal mechanisms for each node X_i in X as $X_i = x_i$ iff $I_{X_i} = \mathrm{do}(x_i)$ (while leaving the causal mechanisms for all other nodes unchanged) i.e., $P_{\mathcal{G}_{\mathrm{do}(X)}}(X)$ is fully determined by the original causal model, the causal mechanisms and $P_{\mathcal{G}_{\mathrm{do}(X)}}(I_X)$ which can be chosen arbitrarily for the exogenous set I_X. Physically, the post-intervention distribution (or the do-conditional) corresponds to additional empirical data that are collected in an experiment, which can, in general, be different from the experiment generating the original, pre-intervention data. For example, when the original experiment involves passive observation of correlations between the smoking tendencies and presence of cancer in a group of individuals, an intervention model may involve forcing certain individuals to take up smoking and then studying their chances of developing cancer. In repeated trials, the proportion of individuals who are passively observed and those that are actively intervened upon may be chosen as desired. The latter type of experiments may not necessarily be ethical but are nevertheless a physical possibility. In certain cases, it may be possible to fully deduce the post-intervention statistics counterfactually from the pre-intervention data (passive observation) alone, and the latter experiment (active intervention) need not be actually performed, sparing us some ethical dilemmas. For example, in a causal structure where all nodes are observed, this is always possible [2]. However, even in simple classical causal structures with unobserved nodes, the post-intervention distribution cannot be completely determined using the pre-intervention distribution alone [2].

d. Further relationships between the pre and post intervention causal models: As explained above, determining the post-intervention distribution from the pre-intervention distribution alone is not possible in the general settings

considered here. However, the compatibility condition of Definition IV.1 along with the defining rules of Equations (4a)-(4d) allows us to derive further useful relationships between these distributions, in particular the three rules of Pearl's do-calculus [2, 51]. These rules have been originally derived in faithful classical causal models satisfying the causal Markov property (1) which does not hold in the general scenarios considered here. Here, we extend these rules to a large class of unfaithful and cyclic non-classical causal models, by noting that the derivation of these rules do not require the Markov property but only the weaker d-separation condition of Definition IV.1 along with the defining rules (4a)-(4d). This is captured in the following theorem and we present a proof in Appendix D 1 for completeness (this is similar to the original proof of [51] but more explicit). In the following, $\mathcal{G}_{\overline{X}}$ denotes the graph obtained by deleting all incoming edges to X and $\mathcal{G}_{\underline{X}}$ denotes the graph obtained by deleting all outgoing edges from a subset X in a graph $\mathcal{G}$, where X is some subset of the observed nodes.

Theorem IV.1. *Given a causal model over a set S of observed nodes, associated causal graph $\mathcal{G}$ and a distribution P_S compatible with $\mathcal{G}$ according to Definition IV.1, the following 3 rules of do-calculus [2] hold for interventions on this causal model.*

- ***Rule 1: Ignoring observations***

$$P_{\mathcal{G}_{\mathrm{do}(X)}}(y|x,z,w) = P_{\mathcal{G}_{\mathrm{do}(X)}}(y|x,w) \qquad \text{if } (Y \perp^d Z|XW)_{\mathcal{G}_{\overline{X}}} \tag{5}$$

- ***Rule 2: Action/observation exchange***

$$P_{\mathcal{G}_{\mathrm{do}(XZ)}}(y|x,z,w) = P_{\mathcal{G}_{\mathrm{do}(X)}}(y|x,z,w) \qquad \text{if } (Y \perp^d Z|XW)_{\mathcal{G}_{\overline{X}\underline{Z}}} \tag{6}$$

- ***Rule 3: Ignoring actions/interventions***

$$P_{\mathcal{G}_{\mathrm{do}(XZ)}}(y|x,z,w) = P_{\mathcal{G}_{\mathrm{do}(X)}}(y|x,w) \qquad \text{if } (Y \perp^d Z|XW)_{\mathcal{G}_{\overline{XZ(W)}}}, \tag{7}$$

where X, Y, Z and W are disjoint subsets of the observed nodes, $Z(W)$ denotes the set of nodes in Z which are not ancestors of W, and the above hold for all values x, y, z and w of X, Y, Z and W.

While the observed distribution in the post-intervention causal model may not be completely specified by the pre-intervention observed distribution alone, considering the underlying causal mechanisms e.g., the states, transformations and measurements involved in the original causal model should allow for the complete specification of the post-intervention distribution. To the best of our knowledge, this problem has not been studied in non-classical and cyclic causal models, we discuss this point in further detail in Appendix C, providing examples of non-classical cyclic causal models where the post-intervention distribution can be calculated from the causal mechanisms. The full solution to this problem will not be relevant to the results of the main paper. Using these concepts, we now define the *affects relation* that is central to the results of this paper.

Definition IV.3 (Affects relation)**.** Consider a causal model associated with a causal graph $\mathcal{G}$ over a set S of observed nodes and an observed distribution P and let X and Y be disjoint subsets of S. If there exists a value x of X such that

$$P_{\mathcal{G}_{\mathrm{do}(X)}}(Y|X=x) \neq P_{\mathcal{G}}(Y),$$

then we say that X *affects* Y.

With this definition, we are ready to state two useful corollaries of Theorem IV.1.

Corollary IV.1. *If X is a subset of observed exogenous nodes of a causal graph $\mathcal{G}$, then for any subset Y of nodes disjoint to X the do-conditional and the regular conditional with respect to X coincide i.e.,*

$$P_{\mathcal{G}_{\mathrm{do}(X)}}(Y|X) = P_{\mathcal{G}}(Y|X).$$

In other words, for any subset X of the observed exogenous nodes, correlation between X and a disjoint set of observed nodes Y in $\mathcal{G}$ guarantees that X affects Y.

Proof. Since X consists only of exogenous nodes, it can only be d-connected to other nodes through outgoing arrows. Then in the graph $\mathcal{G}_{\underline{X}}$ (where all outgoing arrows from X are cut off), X becomes d-separated from all other nodes. This d-separation, $(Y \perp^d X)_{\mathcal{G}_{\underline{X}}}$ implies, by Rule 2 of Theorem IV.1 that $P_{\mathcal{G}_{\mathrm{do}(X)}}(Y|X=x) = P_{\mathcal{G}}(Y|X=x) \quad \forall x$. Further if X and Y are correlated in $\mathcal{G}$, i.e., $\exists x, y$ such that $P_{\mathcal{G}}(y|x) \neq P_{\mathcal{G}}(y)$, the equation previously established along with Definition IV.3 implies that X affects Y. □

Corollary IV.2. *If X and Y are two disjoint subsets of the observed nodes such that $(X \perp^d Y)_{\mathcal{G}_{\mathrm{do}(X)}}$, then X* does not affect *Y and $P_{\mathcal{G}_{\mathrm{do}(X)}}(Y) = P_{\mathcal{G}}(Y)$.*

Proof. The d-separation $(X \perp^d Y)_{\mathcal{G}_{\mathrm{do}(X)}}$ implies the d-separation $(X \perp^d Y)_{\mathcal{G}_{\overline{X}}}$ since $\mathcal{G}_{\mathrm{do}(X)}$ and $\mathcal{G}_{\overline{X}}$ only differ by the inclusion of the intervention nodes I_{X_i} and the corresponding edges $I_{X_i} \longrightarrow X_i$ for each $X_i \in X$. Then by Rule 3 of Theorem IV.1 we have

$$P_{\mathcal{G}_{\mathrm{do}(X)}}(Y|X) = P_{\mathcal{G}}(Y)$$

which by Definition IV.3 implies that X does not affect Y. Further, the d-separation implies the conditional independence $(X \perp\!\!\!\perp Y)_{\mathcal{G}_{\mathrm{do}(X)}}$ i.e.,

$$P_{\mathcal{G}_{\mathrm{do}(X)}}(Y|X) = P_{\mathcal{G}_{\mathrm{do}(X)}}(Y)$$

which along with the result that X does not affect Y yields

$$P_{\mathcal{G}_{\mathrm{do}(X)}}(Y) = P_{\mathcal{G}}(Y).$$ □

Note that X *affects* Y implies that there must be a directed path from X to Y in $\mathcal{G}$ (which is equivalent to X being a cause of Y, cf. Definition II.1). This follows from the contrapositive statement of Corollary IV.2— X affects Y implies that X and Y are not d-separated in $\mathcal{G}_{\mathrm{do}(X)}$ and since this graph has no incoming arrows to X (except those from the intervention nodes in I_X), the only way for X and Y to be d-connected in $\mathcal{G}_{\mathrm{do}(X)}$ is through a directed path from X to Y. However, the converse is not true. A directed path from X to Y in $\mathcal{G}$ does not imply that X affects Y in the presence of fine-tuning (as illustrated in the examples of Appendix A), even though it does imply d-connection between X and Y in $\mathcal{G}_{\mathrm{do}(X)}$ by construction of this graph. This motivates the following classification of the causal arrows $\rightsquigarrow$ between observed nodes. The arrows $\rightsquigarrow$ emanating from or pointing to an unobserved node cannot be operationally probed and hence need not be classified.

Definition IV.4 (Solid and dashed arrows)**.** Given a causal graph $\mathcal{G}$, if two observed nodes X and Y in $\mathcal{G}$ sharing a directed edge $X \rightsquigarrow Y$ are such that X affects Y, then the causal arrow $\rightsquigarrow$ between those nodes is called a *solid arrow*, denoted $X \longrightarrow Y$. Further, all arrows $\rightsquigarrow$ between observed nodes in $\mathcal{G}$ that are *not* solid arrows are called *dashed arrows*, denoted $\dashrightarrow$. In other words, $X \dashrightarrow Y$ for any two RVs X and Y in $\mathcal{G}$ implies that the X does not affect Y.

Remark IV.1 (Exogenous nodes)**.** Note that if X is an exogenous node that is a direct cause of another node Y in a causal graph $\mathcal{G}$ i.e., $X \rightsquigarrow Y$, and X and Y are correlated in the corresponding causal model, then by Corollary IV.1 and Definition IV.4 this would imply that the arrow from X to Y must be a solid one. Applying this to the graphs $\mathcal{G}_{I_X}$ and $\mathcal{G}_{\mathrm{do}(X)}$, where I_X is exogenous and correlated with X by construction (Equations (4a)-(4d)), we can conclude that the arrow from every intervention variable to the corresponding intervened variable must be a solid arrow, i.e., $I_X \longrightarrow X$.

A noteworthy implication that follows from the defining rules is encapsulated in the following lemma.

Lemma IV.2. *Given a causal graph $\mathcal{G}$ and two disjoint subsets X and Y of observed nodes therein,*

$$(X \not\perp\!\!\!\perp Y)_{\mathcal{G}_{\mathrm{do}(X)}} \Rightarrow X \text{ affects } Y.$$

Proof. Suppose that X does not affect Y. By Definition IV.3, this implies that $P_{\mathcal{G}_{\mathrm{do}(X)}}(y|x) = P_{\mathcal{G}}(y)\ \forall x, y$. Further suppose also that $(X \not\perp\!\!\!\perp Y)_{\mathcal{G}_{\mathrm{do}(X)}}$. This means that there exist two distinct values x and x' of X and some value y of Y such that $P_{\mathcal{G}_{\mathrm{do}(X)}}(y|x) \neq P_{\mathcal{G}_{\mathrm{do}(X)}}(y|x')$, which contradicts $P_{\mathcal{G}_{\mathrm{do}(X)}}(y|x) = P_{\mathcal{G}}(y)\ \forall x, y$. Therefore $(X \not\perp\!\!\!\perp Y)_{\mathcal{G}_{\mathrm{do}(X)}}$ must imply X affects Y. □

We note that the affects relation is not transitive in fine-tuned causal models, as illustrated by the following example.

Example IV.1. Consider the causal structure of Figure 5 where all RVs are binary and related by $X = \Lambda$, $Y = W = X \oplus \Lambda$, $Z = Y \oplus W$ with Λ uniformly distributed. Here, both $P_{\mathcal{G}}(Y)$ and $P_{\mathcal{G}}(Z)$ are deterministic distributions. In the graph $\mathcal{G}_{\mathrm{do}(X)}$ obtained by intervening on X, we have $Y = W = X \oplus \Lambda$, $Z = Y \oplus W$ and Λ uniform. Here, since X is not always equal to Λ, $P_{\mathcal{G}_{\mathrm{do}(X)}}(Y|X)$ is no longer deterministic and we have X affects Y, but $P_{\mathcal{G}_{\mathrm{do}(X)}}(Z|X)$ is still the same deterministic distribution irrespective of the value of X since $Y = W$ which implies that X does not affect Z. However, in the graph $\mathcal{G}_{\mathrm{do}(Y)}$, we no longer have $Y = W$ and $P_{\mathcal{G}_{\mathrm{do}(Y)}}(Z|Y)$ is not deterministic, which gives Y affects Z. Therefore affects relations are in general non-transitive in fine-tuned causal models.

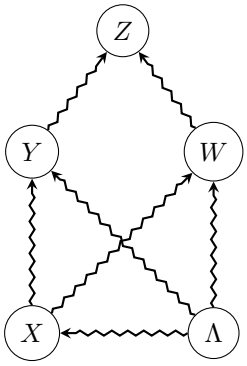


FIG. 5: Causal structure of Example IV.1

C. Conditional and higher-order affects relations

The affects relation defined in Definition IV.3 allows us to consider joint interventions on a subset of the observed nodes S. However certain affects relations where a subset $X \subset S$ that is not a single RV, affects another subset Y, may be "trivial" in the sense that they convey the same information as an affects relation s_X affects Y, where s_X is a proper subset of X, i.e., they can be "reduced" to the latter affects relation. On the other hand, in unfaithful causal models, certain affects relations of the same form can be "non-trivial" in the sense that the information that they convey is not the same as any affects relation from a proper subset of X to Y. To capture this distinction, we introduce higher-order affects relations where we consider whether a set X of RVs affects another disjoint set Y conditioned on an active intervention performed on a third, mutually disjoint subset Z of the RVs. Intuitively these relations are useful because additional interventional information can help us better detect fine-tuned causal influences. More generally, we can also condition on non-interventional information, which leads to the concept of conditional higher-order affects relations. As we will see later in the paper when we bring space-time into the picture, these higher-order affects relations have operational meaning in terms of signalling using joint interventions on space-time random variables, and the conditional higher-order affects relations capture the most general way that agents may signal to each other in our framework. Before we formalise these concepts, some examples would be illustrative.

Example IV.2. Consider a causal model where the only nodes are the observed binary variables X, Y and Z, and the causal graph (Figure 6a) is simply $Z \longrightarrow Y$ and X with no incoming or outgoing arrows. By Definition IV.4 of the solid arrow, Z affects Y and by Corollary IV.2, X does not affect Y. We also have XZ affects Y. This is because $P_{\mathcal{G}_{\mathrm{do}(XZ)}}(Y|XZ) = P_{\mathcal{G}}(Y|XZ)$ and $P_{\mathcal{G}_{\mathrm{do}(Z)}}(Y|Z) = P_{\mathcal{G}}(Y|Z)$ (by exogeneity of X and Z), and using the d-separation condition IV.1) we have $P_{\mathcal{G}}(Y|XZ) = P_{\mathcal{G}}(Y|Z)$. Then Z affects Y implies $P_{\mathcal{G}}(Y|XZ) = P_{\mathcal{G}}(Y|Z) \neq P_{\mathcal{G}}(Y)$ i.e., XZ affects Y. In this example, the node X is entirely superficial as it neither causes nor is a cause of anything else and is therefore completely independent and the affects relation XZ affects Y follows "trivially" from Z affects Y.

Example IV.3. Consider another causal model over the same nodes as the previous example, where the causal graph is a collider from X and Z to Y i.e., $X \rightsquigarrow Y \leftsquigarrow Z$. Furthermore, suppose that Z is uniformly distributed, X is not uniformly distributed and $Y = X \oplus Z$ (where $\oplus$ denotes modulo-2 addition). One can then easily check that the same affects relations as the previous example hold i.e., Z affects Y, X does not affect Y and XZ affects Y, which allows us to classify the causal arrows as in Figure 6b. In this case, Z gives partial information about Y since X is non-uniform, however X and Z taken together give full information about Y. This is in contrast to the previous example where Z as well as XZ gave the same information about Y. More explicitly, the distinguishing condition here is whether or not $P_{\mathcal{G}_{\mathrm{do}(XZ)}}(Y|XZ) = P_{\mathcal{G}_{\mathrm{do}(Z)}}(Y|Z)$; in the previous example this holds, while in the current one it does not.

In general X, Y and Z from the above example may be pairwise disjoint subsets of the observed nodes, and we may condition not only on the set Z (which has been intervened upon), but also on an additional disjoint set of nodes W, upon which an intervention has not been performed. We then have the following definition.

Definition IV.5 (Conditional higher-order affects relation)**.** Consider a causal model associated with a causal graph $\mathcal{G}$ over a set S of observed nodes and an observed distribution P. For four pairwise disjoint subsets X, Y, Z and W of S, we say that X affects Y given $\{\mathrm{do}(Z), W\}$ if there exists values x of X, z of Z and w of W such that

$$P_{\mathcal{G}_{\mathrm{do}(XZ)}}(Y|X=x, Z=z, W=w) \neq P_{\mathcal{G}_{\mathrm{do}(Z)}}(Y|Z=z, W=w). \tag{8}$$

An affects relation X affects Y given $\{\mathrm{do}(Z), W\}$ is a *conditional affects relation* if $W \neq \varnothing$ and an *unconditional affects relation* otherwise. When $Z \neq \varnothing$, it is a *higher-order affects relation*, and a *zeroth-order affects relation* otherwise.

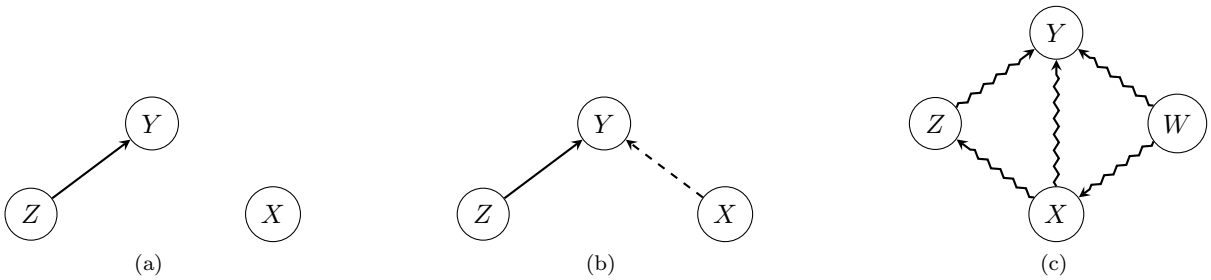


FIG. 6: Causal structures for Examples IV.2, IV.3 and IV.4 respectively.

Definition IV.3 then refers to unconditional zeroth-order affects relations. In general, all of these will be simply called affects relations, unless they need to be explicitly distinguished.

The next lemma (proven in Appendix D 2) establishes the implication of such affects relations for the underlying causal structure.

Lemma IV.3. *For a causal model over a set S of RVs where X, Y, Z and W are any pairwise disjoint subsets of S,*

1. *X affects Y given $do(Z) \Rightarrow X$ is a cause of Y (cf. Definition II.1).*
2. *X affects Y given $\{\mathrm{do}(Z), W\} \Rightarrow X$ is a cause of Y or X is a cause of W.*

It is possible for X not to be a cause of Y and yet satisfy X affects Y given $\{\mathrm{do}(Z), W\}$. A simple example is a 3 node collider causal structure $X \longrightarrow W \longleftarrow Y$ with $W = X.Y$, it is easy to check that X affects Y given W even though X and Y are d-separated. This captures the well known fact that conditioning on a collider can introduce correlations between independent, exogenous variables. Note however that X is a cause of W as implied by the above lemma.

The following lemmas provide useful connections between conditional higher-order and conditional zeroth-order affects relations, their proofs can be found in Appendix D 2. We will often abbreviate higher-order to HO in the following.

Lemma IV.4. *For a causal model over a set S of RVs where X, Y, Z and W are pairwise disjoint subsets of S,*

$$X \text{ affects } Y \text{ given } \{\mathrm{do}(Z), W\} \quad \Rightarrow \quad Z \text{ affects } Y \text{ given } W \quad \text{or} \quad XZ \text{ affects } Y \text{ given } W.$$

Lemma IV.5. *For a causal model over a set S of RVs where X, Y, Z and W are pairwise disjoint subsets of S and X consists only of exogenous nodes,*

$$X \text{ affects } Y \text{ given } \{\mathrm{do}(Z), W\} \quad \Rightarrow \quad XZ \text{ affects } Y \text{ given } W.$$

The converse of Lemma IV.5 is not true, we can have X does not affect Y given $\{\mathrm{do}(Z), W\}$ even when XZ affects Y given W, as we have seen for $W = \varnothing$ in Example IV.2 where X was superficial to the causal model, and the affects relation XZ affects Y trivially followed from the affects relation Z affects Y. Note also that the implication of the above lemma does not hold in general when X is not exogenous. This is because in fine-tuned causal models (rather counter-intuitively), Z affects Y does not imply that any set of RVs containing Z also affects Y, which was a step required in the above proof. The following example illustrates this.

Example IV.4. Consider the causal structure of Figure 6c. Suppose that the exogenous W is uniformly distributed and the variables are related as $Y = X \oplus Z \oplus W$, $Z = X$, $X = W$. This gives $Y = X = Z = W$ and hence $P_{\mathcal{G}}(Y) = P_{\mathcal{G}}(W)$ is uniform. In the graph $\mathcal{G}_{\mathrm{do}(Z)}$, we have $Y = X \oplus Z \oplus W$, and $X = W$ which gives $Y = Z$ and hence $P_{\mathcal{G}_{\mathrm{do}(Z)}}(Y|Z)$ is deterministic. This gives Z affects Y. In the graph $\mathcal{G}_{\mathrm{do}(XZ)}$, we only have the relation $Y = X \oplus Z \oplus W$ which implies that $P_{\mathcal{G}_{\mathrm{do}(XZ)}}(Y|XZ)$ is uniform and hence that XZ does not affect Y. Note that we also have X affects Y given $\mathrm{do}(Z)$.

Definition IV.5 does not yet fully capture the notion of "reducibility" or "triviality" of certain affects relations. consider Example IV.3 again and add a superficial observed node V with no incoming or outgoing arrows. Then we have both the higher-order affects relations X affects Y given $\mathrm{do}(Z)$ and XV affects Y given $\mathrm{do}(Z)$. However, the

addition of V adds no information to the original affects relation since $P_{\mathcal{G}_{\mathrm{do}(XZV)}}(Y|XZV) = P_{\mathcal{G}_{\mathrm{do}(XZ)}}(Y|XZ)$ (i.e., V does not affect Y given $\mathrm{do}(XZ)$). In other words, the affects relation XV affects Y given $\mathrm{do}(Z)$ is reducible to the affects relation X affects Y given $\mathrm{do}(Z)$. Based on this idea, we propose the following criterion for distinguishing between reducible and irreducible affects relations.

Definition IV.6 (Reducible and irreducible affects relations)**.** For a causal model defined over a set S of observed nodes, the affects relation X affects Y given $\{\mathrm{do}(Z), W\}$ between pairwise disjoint subsets X, Y, Z and W of S is said to be *reducible* if there exists a proper subset s_X of X such that s_X does not affect Y given $\{\mathrm{do}(Z\tilde{s}_X), W\}$, where $\tilde{s_X} := X \backslash s_X$. Conversely, if for all proper subsets s_X of X, s_X affects Y given $\{\mathrm{do}(Z\tilde{s}_X), W\}$, the affects relation X affects Y given $\{\mathrm{do}(Z), W\}$ is said to be *irreducible*.

Then we have the following lemmas, which make clear why the above definition captures a notion of "reduction" of the affects relation. Proofs of these lemmas can be found in Appendix D 2.

Lemma IV.6. *For every reducible affects relation X affects Y given $\{\mathrm{do}(Z), W\}$, there exists a proper subset $\tilde{s}_X$ of X such that $\tilde{s}_X$ affects Y given $\{\mathrm{do}(Z), W\}$.*

Lemma IV.7. *For a causal model over a set S of RVs of which X_1, X_2, Y, Z and W are pairwise disjoint subsets,*

$$X_1 \text{ affects } Y \text{ given } \{\mathrm{do}(Z), W\} \text{ and } X_2 \text{ does not affect } Y \text{ given } \{\mathrm{do}(ZX_1), W\}$$

$$\Downarrow$$

$$X_1X_2 \text{ affects } Y \text{ given } \{\mathrm{do}(Z), W\}.$$

Definition IV.6 classifies the relation XZ affects Y as reducible in Example IV.2 (Fig. 6a), and irreducible in Example IV.3 (Fig. 6b). Note that checking for the (ir)reducibility of an affects relation involves considering an affects relation of a greater order than the original one, where the order of X affects Y given $\{\mathrm{do}(Z), W\}$ is measured by the cardinality $|Z|$ of Z.

The following lemma (proven in Appendix D 2) relates conditional affects relations to unconditional affects relations such that the irreducibility of the former implies the irreducibility of the latter. As we will later see, this will allow us to restrict to unconditional affects relations without loss of generality when considering their space-time embeddings (cf. Remark V.2).

Lemma IV.8. *For a causal model over a set S of RVs where X, Y, Z and W are pairwise disjoint subsets of S,*

1. *X affects Y given $\{\mathrm{do}(Z), W\} \Rightarrow$ X affects YW given $do(Z)$.*
2. *X affects Y given $\{\mathrm{do}(Z), W\}$ is irreducible $\Rightarrow$ X affects YW given $do(Z)$ is irreducible.*
3. *X affects YW given $do(Z) \Leftrightarrow$ X affects Y given $\{\mathrm{do}(Z), W\}$ or X affects W given $do(Z)$.*

The converse does not hold for the first two statements of this lemma, as illustrated by the following counter-examples. For part 1, consider again Example IV.3 with the superficial observed node V having no incoming or outgoing arrows. Here we have Z affects Y and Z affects VY and yet Z does not affect V given Y (Z, V and Y play the role of X, Y and Z in the above lemma with $W = \varnothing$). For part 2, consider the causal structure $X_1 \dashrightarrow W \dashleftarrow X_2 \longrightarrow Y$ with all variables binary, $W = X_1 \oplus X_2$, $Y = X_2$, X_1 and X_2 uniformly distributed. Taking $X = X_1X_2$, it is easy to verify that we have X affects Y given W, X affects YW and it is irreducible, while X affects Y given W is reducible to X_2 affects Y given W.

Using this, we obtain a stronger version of Lemma IV.3 as a corollary of Lemmas IV.3 and IV.8 (see Appendix D 2 for a proof).

Corollary IV.3. *For a causal model over a set S of RVs where X, Y, Z and W are any pairwise disjoint subsets of S,*

1. *X affects Y given $do(Z)$ is irreducible $\Rightarrow$ for each element $e_X \in X$ there exists an element $e_Y \in Y$ such that e_X is a cause of e_Y.*
2. *X affects Y given $\{\mathrm{do}(Z), W\}$ is irreducible $\Rightarrow$ for each element $e_X \in X$ there exists an element $e_{YW} \in YW$ such that e_X is a cause of e_{YW}.*

Remark IV.2. Note that in the language of conditional HO affects relations, Pearl's 3rd rule of do-calculus (Theorem IV.1) can be written in the equivalent form

$$(Y \perp^d Z|XW)_{\mathcal{G}_{\mathrm{do}(XZ(W))}} \quad \Rightarrow \quad X \text{ does not affect } Y \text{ given } \{\mathrm{do}(Z), W\},$$

where $Z(W)$ is the set of nodes in Z that are not ancestors of W.

Remark IV.3. We have seen in Definition IV.4 that a dashed arrow from X to Y corresponds to causation in the absence of the corresponding zeroth-order affects relation X affects Y. A natural question to ask is whether all dashed arrows in a causal model can be detected using higher-order affects relations. If we consider causal models with no latent nodes, then this is the case. Such a model is entirely classical and the causal mechanisms consist of functional equations i.e., for each node Y, a function f_Y taking as input the parent variables $\mathrm{par}(Y)$ and an independent, exogenous error variable E_Y that completely determines Y as $Y = f_Y(\mathrm{par}(Y), E_Y)$. The meaning of saying that X is a parent of Y is that f_Y has a nontrivial dependence on the input X, i.e., there exists a fixed value of all other inputs of f_Y such that changing the value of X produces a change in the function value. This is precisely captured by the higher-order affects relation X affects Y given $\mathrm{do}(\mathrm{par}(Y) \setminus X, E_Y)$. Therefore, given any unfaithful causal model where all nodes, including the error nodes are observed and can be intervened upon, full causal discovery is possible i.e., whether there exists a causal link $X \rightsquigarrow Y$ between any two nodes X and Y in the model, and whether this is a dashed or solid arrow can be determined by interventions in this case. While requiring all the nodes to be observable might be quite a strong assumption, we are not aware of a method for full causal discovery of arbitrary unfaithful causal models in previous literature even under this assumption. By introducing the new concept of higher-order affects relations, our framework suggests an advantage for the classical causal discovery problem for unfaithful causal models. The further exploration of the connections between our framework and the general causal discovery problem is left to future work.

D. Relationships between concepts

Due to the presence of fine-tuning and the introduction of the 2 types of causal arrows (solid and dashed), a number of concepts that are equivalent in faithful causal models are not equivalent for the causal models described in our framework. We summarise some of the relationships between the concepts arising in our causal modelling framework, before bringing space-time structure into the picture. This subsection can be skipped at the first reading.

The relationships are illustrated in Figure 7. The reason for every implication is explained in the figure caption, and for every implication that fails, we provide a counter-example below. There are 14 implications in Figure 7 that do not hold. Some of these can be explained by the same counter-example or are immediately evident from the definitions. Therefore we first group these 14 cases based on the corresponding counter-example or argument needed for explaining them, in the end we will only need a few distinct counter-examples to cover all these cases. Note that if we restrict to faithful and/or acyclic causal models, not all of these non-implications would hold. For instance, in the case of faithful and acyclic causal models commonly considered in the literature, non-implications 1, 2, 3, 4, 5, 9 and 12 will become implications. This section does not cover all implication or non-implications found in this paper, since some of these also involve the newly introduced conditional HO affects relations. For this, we refer the reader to the previous sections. Here we consider relationships between certain basic notions such as correlation vs causation vs affects relations (unconditional zeroth-order ones), to illustrate how these differ in the fine-tuned case.

1. **Non-implication 1:** In unfaithful causal models, X and Y can be independent even when they are d-connected, as we have seen in the examples of Figure 2.

2. **Non-implications 2, 11, 18:** These are covered by Example IV.5.

3. **Non-implications 3, 6, 8, 13:** These are covered by Example IV.6.

4. **Non-implications 4, 5:** X is a cause of Y does not imply that it is a direct cause of Y, it can be an indirect cause. Further X can affect Y even when it is an indirect cause, for example $X \longrightarrow Z \longrightarrow Y$.

5. **Non-implication 7:** This is covered by Example IV.7.

6. **Non-implication 9:** It is evident that "X is a direct cause of Y" does not imply $X \dashrightarrow Y$, since it can also be a cause through a solid arrow.

7. **Non-implications 10, 12:** These are just a consequence of the fact that correlation does not imply causation. Correlation between X and Y can arise when they share a common cause, without being a cause (direct or indirect) of each other.

8. **Non-implications 14, 17:** In a simple common cause scenario, i.e., $Z \longrightarrow X$ and $Z \longrightarrow Y$ with $X = Y = Z$, X does not affect Y however X is correlated with Y and there is no dashed arrow from X to Y.

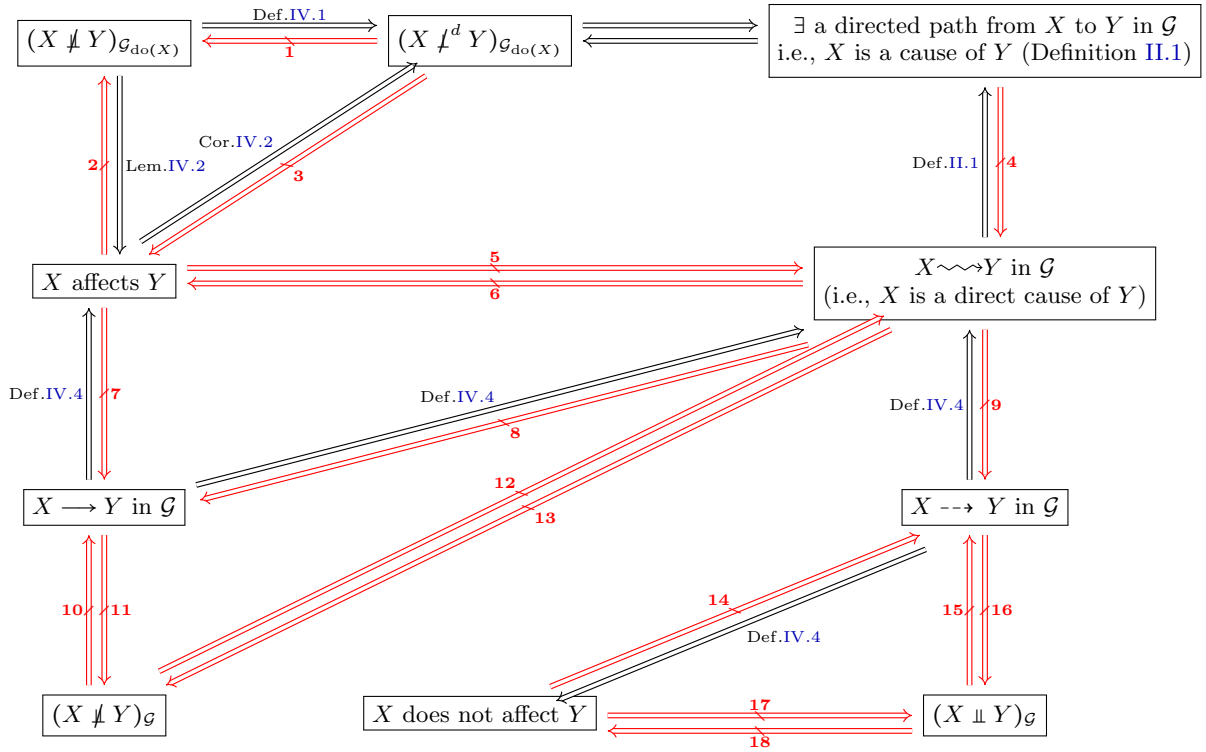


FIG. 7: **Relationships between concepts relating to causal models:** The black arrows denote implications while red (crossed out) arrows denote non-implications. The numbers label the counter-examples corresponding to each non-implication, which are explained in the main text. The equivalence between "∃ a directed path from X to Y in $\mathcal{G}$" and $(X \not\perp\!\!\!\perp^{d} Y)_{\mathcal{G}_{\mathrm{do}(X)}}$ is explained in the paragraph following Corollary IV.2. $X \longrightarrow Y$ and $X \dashrightarrow Y$ imply $X \rightsquigarrow Y$ since solid and dashed arrows are simply special instances of the more general, squiggly arrow by Definition IV.4. This graph is complete in the sense that, given any ordered pair of statements (ϕ_1, ϕ_2) from the 10 that form the vertices of this graph, one can deduce whether or not $\phi_1 \Rightarrow \phi_2$ as follows: if there exists a directed path from ϕ_1 to ϕ_2 that consists only of the implication arrows (black), then $\phi_1 \Rightarrow \phi_2$ and otherwise, $\phi_1 \not\Rightarrow \phi_2$.

9. **Non-implication 15:** It is evident that independence of X and Y does not imply that there is a dashed arrow between them, they can also be d-separated.

10. **Non-implication 16:** This is covered by Example IV.8.

Example IV.5. Consider the causal structure of Figure 8. Let the three variables S, E and H be binary and correlated as $H = S \oplus E$ and $S = E$. These relations imply that $H = 0$ deterministically while $S = E$. Now, when we intervene on E, we can choose its value independently of S and whenever we choose $E \neq S$, we will see that $H = 1$ occurs with non-zero probability. In other words, there exists a value e of E such that $P(H = 1|\mathrm{do}(e)) \neq P(H = 1) = 0$ i.e., E affects H. As E is a direct cause of H in $\mathcal{G}$, this further implies that the causal arrow from E to H is a solid one, even though E and H are independent in both the pre and post-intervention causal models i.e., $(E \perp\!\!\!\perp H)_{\mathcal{G}}$ and $(E \perp\!\!\!\perp H)_{\mathcal{G}_{\mathrm{do}(X)}}$ both hold, the former since H is deterministic in the original causal model, irrespective of the value of E and the latter since H is uniform in the post-intervention model, again irrespective of the value of E. Therefore the existence of an affects relation between two sets of observed variables does not imply correlation between them either in the pre or the post intervention causal model. Further, S does not affect H since the exogeneity of S implies that $P_{\mathcal{G}_{\mathrm{do}(S)}}(H|S) = P_{\mathcal{G}}(H|S)$ (Corollary IV.1), and the independence of S and H in $\mathcal{G}$ gives $P_{\mathcal{G}}(H|S) = P_{\mathcal{G}}(H)$.

Example IV.6 (Jamming)**.** Consider the causal structure of Figure 9a where $B \dashrightarrow A$, $B \dashrightarrow C$ and the RVs A and C share an unobserved common cause Λ. By Definition IV.4 of the dashed arrows, we have B does not affect A and B does not affect C. Suppose that B affects the set AC. When A, B and C are binary, a probability distribution compatible with this situation is one where all 3 RVs are uniformly distributed and correlated as $B = A \oplus C$, where $\oplus$ stands for modulo-2 addition. Then, A and C individually carry no information about B but A and C jointly

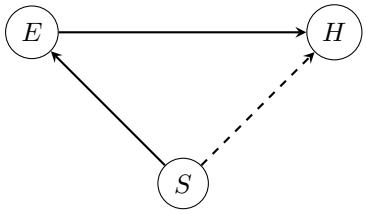


FIG. 8: **Affects relation does not imply correlation:** This is a causal structure for Example IV.5 which demonstrates a scenario where E affects H even though $P_{EH} = P_E P_H$, i.e., solid arrows can also be fine-tuned and the ability to detect causation through an active intervention does not imply that we will see correlation upon passive observation.

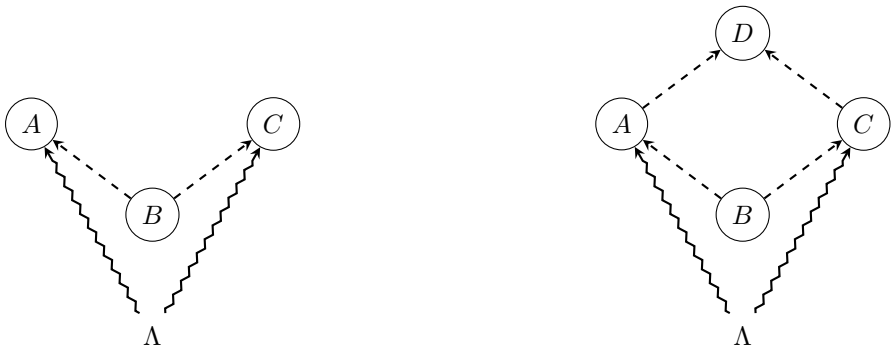


FIG. 9: **Some fine-tuned causal structures:** (a) The jamming causal structure of Example IV.6. Note that the common cause Λ is essential to this example, because without Λ, A and C would be d-separated given B which would imply the conditional independence $P_{AC|B} = P_{A|B}P_{C|B}$. The dashed arrows would imply the independence of A and B as well as C and B and hence the observed distribution would factorise as $P_{ABC} = P_A P_B P_C$. Then no pairs of disjoint subsets of $\{A, B, C\}$ would affect each other contrary to the original example. (b) Causal structure for Example IV.7 where B affects D even though there is no solid arrow path from B to D.

determine the exact value of B. In this case, B is a cause of A and of C but, due to fine-tuning, B and A are uncorrelated, as are B and C, and there are no pairwise affects relations. This means that the causal influence of B on A (or B on C) can only be detected when A, B and C are jointly accessed. The common cause is crucial to this example as explained in Figure 9a, and the causal structure compatible with the distribution and affects relations of this example is not unique. An alternative causal structure that is compatible with correlations and affects relations of this example is where one of the dashed arrows $B \dashrightarrow A$ or $B \dashrightarrow C$ is dropped.

This example by itself makes no reference to space-time or the tripartite Bell scenario. However, if the variables A, B and C are embedded in a pairwise space-like separated way and taken to correspond to the output of Alice, input of Bob and output of Charlie respectively, this becomes a special case of the tripartite jamming scenario of [30, 31] (Figure 3).[9] In the rest of the paper, such examples, where an RV has dashed arrows to a set of RVs will be referred to as instances of "jamming" in accordance with the terminology of [30], irrespective of the space-time configuration. We will further discuss the relation of such causal models to space-time structure later in the paper.

Example IV.7. Consider a causal model over observed variables $\{A, B, C, D\}$ associated with the causal graph $\mathcal{G}$ given in Figure 9b. Here, there are no pairs of variables sharing an edge such that one of them affects the other. A correlation compatible with this graph is obtained by taking $B = A \oplus C = D$ where all variables are binary and uniformly distributed. Here, B affects D even though there are no solid arrow paths from B to D.

[9] Barring the slight change of notation: In Figure 3, A and C correspond to the inputs of Alice and Charlie while X and Z correspond to the outputs that are jammed by B. We do not make a distinction between inputs and outputs in general since we will also consider situations where the jamming variable is not exogenous for example.

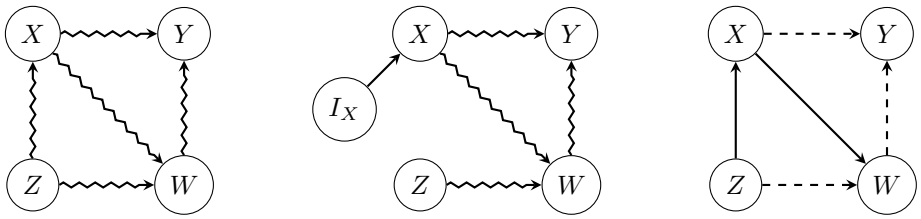


FIG. 10: **Dashed arrow (or non-affects relation) does not imply independence:** (a) The original causal structure $\mathcal{G}$ of Example IV.8, before the causal arrows are classified according to Definition IV.4. (b) The corresponding causal structure $\mathcal{G}_{\mathrm{do}(X)}$ when the node X is intervened upon. (c) The causal structure $\mathcal{G}$ after all the arrows have been classified as explained in the main text. The example shows that even though we have $X \dashrightarrow Y$ in $\mathcal{G}$, there exists a causal model compatible with this graph such that X and Y are correlated in $\mathcal{G}$ in this causal model.

Example IV.8. Consider the causal structure of Figure 10a with the variables X, Y, W and Z taken to be binary. Suppose the causal mechanisms of the model are $X = Z$, $W = X \oplus Z$ and $Y = X \oplus W$ with the exogenous variable Z being uniformly distributed. This reduces to $W = 0$ (deterministically) and $Y = X = Z$. Since Z is uniformly distributed, $P_{\mathcal{G}}(Y)$ is also uniform and since X and Y are perfectly correlated in $\mathcal{G}$, $P_{\mathcal{G}}(Y|X)$ is deterministic. Now consider the graph $\mathcal{G}_{\mathrm{do}(X)}$ shown in Figure 10b. The causal mechanism for X here is fully specified by the distribution over P_{I_X} which can be arbitrary. For the remaining variables we have $W = X \oplus Z$, $Y = X \oplus W$ and Z is uniformly distributed, which gives $Y = Z$. The d-separation $(Z \perp^d X)_{\mathcal{G}_{\mathrm{do}(X)}}$ implies the independence of Z and X in $\mathcal{G}_{\mathrm{do}(X)}$ and hence the independence of Y and X in $\mathcal{G}_{\mathrm{do}(X)}$ and since Z is uniformly distributed here, so is Y i.e., $P_{\mathcal{G}_{\mathrm{do}(X)}}(Y|X) = P_{\mathcal{G}_{\mathrm{do}(X)}}(Y)$ and both equal the uniform distribution. From before, we had noted that $P_{\mathcal{G}}(Y)$ is also uniform, which gives $P_{\mathcal{G}_{\mathrm{do}(X)}}(Y|X) = P_{\mathcal{G}}(Y)$ or X *does not affect* Y. Therefore, by definition IV.4, the causal arrow from X to Y must be a dashed one, even though we have seen that $(X \not\perp Y)_{\mathcal{G}}$.

The remaining causal arrows of Figure 10a can also be classified into solid and dashed arrows as done for $X \rightsquigarrow Y$ in the above example. For example, X *affects* W can be established by noting that $P_{\mathcal{G}_{\mathrm{do}(X)}}(W|X)$ is uniform (since $W = X \oplus Z$ with X and Z independent in $\mathcal{G}_{\mathrm{do}(X)}$ and Z is uniform) while $P_{\mathcal{G}}(W)$ is deterministic. Therefore we have $X \longrightarrow W$ in $\mathcal{G}$. Similarly, W *does not affect* Y, Z *affects* X and Z *does not affect* W can also be established and we obtain the graph of Figure 10c as the original causal structure $\mathcal{G}$ once all the arrows of Figure 10a have been classified.

Further examples can be found in Appendix A where we discuss how conditional independences and affects relations can be deduced from the causal model in our framework.

V. THE FRAMEWORK, PART 2: SPACE-TIME

We now turn to space-time structure and the relevant concepts needed for studying its relation to causality. Section V A introduces our way of modelling space-time structure and the concept of *ordered random variables* (definition V.1). In Section V B, we define what it means to embed a causal model in a space-time structure (Definition V.6). We then characterise in Section V C, what it means for a causal model to be compatible with an embedding in a space-time (Definition V.7), which formalises the requirement that signalling outside the space-time future is not possible using the affects relations of the embedded causal model. Finally, in Theorem V.1 of Section V D, we provide necessary and sufficient conditions for compatibility.

A. Space-time structure

We model space-time simply by a partially ordered set $\mathcal{T}$ without assuming any further structure/symmetries. A particular example of $\mathcal{T}$ is Minkowski space-time, where the partial order corresponds to the light-cone structure and the elements of $\mathcal{T}$ can be seen space-time coordinates in some frame of reference. Our results will only depend on the order relations of $\mathcal{T}$ and not on the representation of its particular elements. To make operational statements about $\mathcal{T}$, we must embed physical systems into it. In our case, we can only do so for the observed systems in the causal model which are random variables. We embed them in this space-time by assigning an element of $\mathcal{T}$ to each random variable which then specifies its space-time location (thereby producing an *ordered random variable* or ORV),

and assigning a subset of $\mathcal{T}$ to each ORV which specifies the locations in the space-time at which the ORV can be "accessed". Here, the order of an ORV corresponds to that of the space-time $\mathcal{T}$ (and not of the causal model) i.e., ORVs can be seen as abstract versions of space-time random variables.

Definition V.1 (Ordered random variable (ORV))**.** Given a RV X, we can assign to it a location $O(X) \in \mathcal{T}$. An ORV $\mathcal{X}$ is then the pair $\mathcal{X} := (X, O(X))$. We can extend the definition of O to ORVs, so that $O(\mathcal{X})$ is interpreted to mean $O(X)$.

We use $\prec$, $\succ$ and $\nprec\nsucc$ to denote the order relations for a given partially ordered set $\mathcal{T}$, where for $\alpha, \beta \in \mathcal{T}$, $\alpha \nprec\nsucc \beta$ corresponds to α and β being unordered with respect to $\mathcal{T}$. This is different from $\alpha = \beta$ which corresponds to the two elements being equal. These relations carry forth in an obvious way to ORVs and we say for example that 2 ORVs $\mathcal{X}$ and $\mathcal{Y}$ are ordered as $\mathcal{X} \prec \mathcal{Y}$ iff $O(\mathcal{X}) \prec O(\mathcal{Y})$. Note however that when we write $\mathcal{X} = \mathcal{Y}$ we mean $X = Y$ and $O(\mathcal{X}) = O(\mathcal{Y})$.

Definition V.2. The *inclusive future* of an ORV is the set

$$\overline{\mathcal{F}}(\mathcal{X}) := \{\alpha \in \mathcal{T} : \alpha \succeq O(\mathcal{X})\}.$$

Note that $\mathcal{X} \in \overline{\mathcal{F}}(\mathcal{X})$ but $\mathcal{X} \notin \mathcal{F}(\mathcal{X})$, hence the name "inclusive" future. Then, we say that an ORV $\mathcal{Y}$ lies in the inclusive future of an ORV $\mathcal{X}$ iff $O(\mathcal{Y}) \in \overline{\mathcal{F}}(\mathcal{X})$. In a slight abuse of notation, we will simply write this as $\mathcal{Y} \in \overline{\mathcal{F}}(\mathcal{X})$, which is equivalent to $\mathcal{X} \preceq \mathcal{Y}$. Further, any probabilities written in terms of ORVs should be understood as being probability distributions over the corresponding random variables. In the rest of the paper, whenever we use the term "future", this should be understood as inclusive future.

Remark V.1. When considering causal loops or closed timelike curves (CTC)[10], one typically imagines a cyclic space time whose light cone structure is not a partial order, but a pre-order. This is the case in general relativity where the space-time structure implies a causal structure and having a CTC is a property of the space-time. Here, we have separated causality from space-time such that causal loops are a property of the causal model (see Section VI), and any causal loop embedded in a space-time (partial or pre-ordered) as described in the following section would form a CTC. We will consider how such cyclic causal models can be compatibly embedded in a space-time i.e., without leading to signalling outside the future, and the more interesting case is when we take a partially ordered space-time such as Minkowski space-time. Through this approach, we will see that it is possible to have a CTC in Minkowski space-time that does not lead to superluminal signalling, since it is possible for the signalling properties of a causal model to respect the partial order even while the causal relations are cyclic. The problem would be in a sense trivial if the space-time is also a preorder, since for any cyclic causal structure (which defines a pre-order relation), one can always find a corresponding pre-ordered space-time that compatibly embeds it.

B. Embedding of a causal model in a space-time structure

We have discussed two types of order relations: the pre-order encoded by the arrows $\rightsquigarrow$ of the causal structure, and the partial order specified by the order relation $\prec$ of the partial order $\mathcal{T}$. These are two distinct concepts, and within our framework can be set independently of one another. We first formalise how a given causal model may be embedded in a space-time structure, and in the next section, we introduce a compatibility condition that connects the two that aims to capture when a causal structure can be embedded in the partial order $\mathcal{T}$. This compatibility condition is based on the idea of ensuring that it is impossible to signal outside the future as encoded by the partial order $\mathcal{T}$[11]. Whether signalling is possible depends on where random variables can be accessed, and so we first introduce the concept of an accessible region, which is the subset of $\mathcal{T}$ at which it is possible to have a copy of a random variable. Since we are dealing with classical random variables, it makes sense to imagine these being broadcast, i.e., sending a copy to all points in the accessible region.

Definition V.3 (Copy of a RV)**.** Consider a causal model over a set of observed variables S. A RV $X' \in S$ is a *copy* of $X \in S$ if the only parent of X' is X, and if $X' = X$. It is often convenient to think of copying a random variable X in the causal model, where the copy is not initially included in the model. To do so, we augment the causal graph

[10] By CTC we mean any situation in which a causal model whose causal structure has a loop is embedded in space-time (cf. Definition V.6). This leads to causal influences in both directions between two points in the space-time.

[11] It may be helpful to think of $\mathcal{T}$ as a Minkowski space-time, with the partial order specified by the light-cone structure.

with a new node X' whose only parent is the node X and such that $X' = X$ (the graph has $X \longrightarrow X'$ added). We usually do not draw the augmented causal model, but instead keep the copies implicit. We also extend the definition of a copy to ordered random variables so that $\mathcal{X}'$ is a copy of $\mathcal{X}$ whenever the corresponding RV X' is a copy of the RV X.

Note that each RV affects each of its copies. We can then define the accessible region of a RV to be the region of $\mathcal{T}$ in which it is possible to have a copy of the RV. In essence, we can imagine each RV being copied throughout its accessible region.

Definition V.4 (Accessible region of a RV/ORV)**.** Given a causal model over a set of observed variables S, and a partial order $\mathcal{T}$, for each random variable $X \in S$ we can define an *accessible region* $\mathcal{R}_X \subseteq \mathcal{T}$ intended to represent the set of points in $\mathcal{T}$ at which it is possible to have a copy of X. The *inaccessible region* of X is then the complement $\tilde{\mathcal{R}}_X = \mathcal{T} \setminus \mathcal{R}_X$ and represents the set of points at which it is impossible to have a copy of X. We can naturally extend this definition to ORVs by taking the accessible region of an ORV $\mathcal{X} = (X, O(X))$ to be the accessible region of X.

We also want a notion of accessible region for sets of RVs/ORVs. The accessible region of a set can be thought of as the locations at which there can be a copy of all of the random variables in the set. This motivates taking the intersection of the accessible regions of the individual elements, since if the accessible region of the set were any larger than this, it would contradict the definition of accessible region for at least one individual element of the set.

Definition V.5 (Accessible region of a set of RVs/ORVs)**.** Given a set $S = \{S_i\}_i$ of RVs we define the accessible region of S by $\mathcal{R}_S = \bigcap_{S_i \in S} \mathcal{R}_{S_i}$. For the empty set $\varnothing$, the accessible region is defined to be $\mathcal{R}_\varnothing := \mathcal{T}$.

Definition V.6 (Embedding)**.** Given a set of RVs S, an *embedding of* S in a partially ordered set $\mathcal{T}$ produces a corresponding set of ORVs $\mathcal{S}$ by assigning a location $O(X) \in \mathcal{T}$, and an accessible region $\mathcal{R}_X$ to each RV X, such that the associated ORV is $\mathcal{X} = (X, O(X))$. An embedding of a set of RVs is called *non-trivial* if no two RVs X and Y such that X affects Y are assigned the same location in $\mathcal{T}$.

The set of RVs S we will wish to embed will typically be related by a causal model or a set of affects relations. We have seen that when analysing affects relations, it is useful to augment the original causal model with an additional set of RVs corresponding to the intervention nodes. In the following, whenever we refer to an embedding of a causal model or a set of affects relations in a partial order, this must be understood as an embedding of the original set of RVs S associated with causal model/affects relations, the non-triviality of the embedding will also only concern the embedding of the original set of RVs S. For simplicitly, we will assume that every hypothetical intervention node I_X that may be introduced to model interventions on an RV $X \in S$ is embedded at the same location as X (even though I_X affects X by construction). Our results are not affected by this assumption, it is a mere simplification.

C. Compatibility of a causal model with an embedding in space-time

Up to here there are no conditions on how the locations and accessible regions are set—in particular, these need not be related with the notion of future defined on $\mathcal{T}$. We now introduce a compatibility condition that connects these concepts together, which aims to capture the intuition that signalling outside the (inclusive) future should not be possible. As this intuition is quite non-trivial to formalise for general, unfaithful causal models, we will first motivate the important aspects of the definition with examples, before formally stating it. For this, we will first consider the case of faithful causal models, then unfaithful causal models with interventions only on single nodes and finally the general case of unfaithful causal models with joint interventions. For all the examples in the following paragraphs we will take $\mathcal{T}$ to be Minkowski space-time and embed RVs such that the accessible region of each RV coincides with its inclusive future.

a. Compatibility for faithful causal models: For faithful causal models, if X and Y are 2 RVs, X is a cause of Y in a causal structure $\mathcal{G}$ i.e., $X \rightsquigarrow \ldots \rightsquigarrow Y$ in $\mathcal{G}$ is equivalent to X affects Y. Therefore, if we demand that whenever X affects Y for any two RVs X and Y in the model, Y must be embedded in the future of X in the space-time, this ensures that all causal influences propagate from past to future and consequently that there is no signalling outside the future for the given embedding of the model.[12]

[12] Note that such an embedding is always possible for acyclic causal models but impossible for causal models with certain types of causal loops (Lemma VI.2) and possible for causal models with certain other types of causal loops as we will show in Section VI.

b. Compatibility for unfaithful causal models with single node interventions: The above condition for faithful models is insufficient to rule out such signalling in unfaithful models since affects and cause become inequivalent notions here, and we must also consider affects relations involving sets of RVs. For example, in the jamming causal structure (Example IV.6), if we embed A and C outside the future of B, but such that there are points in the intersection of the futures of A and C that are also outside the future of B, then signalling is possible. We first consider affects relations of the form X affects S where X is an RV and S is a set of RVs. Operationally, this means that given access to a copy of all elements of S, one can learn information about the intervention performed on X. Then, in order to avoid signalling outside the future by means of the affects relation X affects S, a necessary and sufficient condition on the embedding would be to take the accessible regions to coincide with the inclusive futures and $\mathcal{R}_S \subseteq \mathcal{R}_X$, which would ensure that the joint future of all elements in S is contained in the future of X. Note that this does not imply that all causal influences (which may be hidden due to fine-tuning) must propagate from past to future, only that any observable signal propagates from past to future (e.g., the jamming scenario of Figure 3).

c. Compatibility for unfaithful causal models with multi-node interventions: Consider a general affects relation of the form S_1 affects S_2 for two disjoint subsets S_1 and S_2 of RVs (possibly arising from an unfaithful causal model). If in analogy to the previous case, we demand that any compatible embedding must be such that $\mathcal{R}_{S_2} \subseteq \mathcal{R}_{S_1}$ with all accessible regions coinciding with the corresponding inclusive futures, this would be too restrictive in the present case. Take the simple Example IV.2 where $Z \longrightarrow Y$ and X is an isolated node with no in or out edges. Then clearly XZ affects Y but we would only require Y to be in the future of Z and not also in the future of X (which trivially affects it given Z). On the other hand, in the causal structure of Example IV.3, Y depends on both the exogenous nodes X and Z and we would expect that Y must be embedded in the joint future of X and Z to avoid signalling outside the future. To establish that embedding Y in the joint future of X and Z is necessary in the latter case and not the former and to avoid imposing too strong constraints on the embedding, we must also consider the higher-order affects relation X affects Y given $\mathrm{do}(Z)$.

d. Operational meaning of a higher-order affects relation: Operationally, the conditional HO affects relation X affects Y given $\{\mathrm{do}(Z), W\}$ means that an agent Alice who can intervene on X can signal to an agent Bob having access to Y if Bob also has access to information about interventions performed on some set Z along with information about some other set W (upon which an intervention was not performed). If the RVs in these sets are embedded in a space-time, in order for the affects relation X affects Y given $\{\mathrm{do}(Z), W\}$ to not lead to signalling outside the space-time future, we must embed the RVs such that the joint future of Y, Z and W (i.e., where they are jointly accessible by Bob) is contained in the future of X.

Furthermore, a given HO affects relation, X affects Y given $\{\mathrm{do}(Z), W\}$ may itself contain some redundancies if X is a set of RVs (as we have seen in Example IV.3), such that it can be reduced to the HO affects relation $\tilde{s}_X$ affects Y given $\{\mathrm{do}(Z), W\}$ for some proper subset $\tilde{s}_X$ of X (Lemma IV.6). In such cases we only need to impose that the joint future (or joint accessible region) of Y and Z is contained in that of the smaller set $\tilde{s}_X$.

The following definition based on this intuition allows us to decide when a set of affects relations can be compatibility embedded in a space-time.

Definition V.7 (Compatibility of a set of affects relations with an embedding in a partial order (**compat**))**.** Let $\mathcal{S}$ be a set of ORVs formed by embedding a set of RVs S in a partially ordered set $\mathcal{T}$ with embedding $\mathscr{E}$. Then a set of affects relations $\mathscr{A}$ is said to be *compatible* with the embedding $\mathscr{E}$ if the following conditions hold:

- **compat1:** Let $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathcal{S}$ be disjoint non-empty subsets of ORVs, and $\mathcal{S}_3, \mathcal{S}_4$ be two more subsets (possibly empty) disjoint from each other and $\mathcal{S}_1$ and $\mathcal{S}_2$. If $\mathcal{S}_1$ affects $\mathcal{S}_2$ given $\{\mathrm{do}(\mathcal{S}_3), \mathcal{S}_4\}$ is in $\mathscr{A}$ and is irreducible with respect to the affects relations in $\mathscr{A}$, then $\mathcal{R}_{\mathcal{S}_2\mathcal{S}_3\mathcal{S}_4} = \mathcal{R}_{\mathcal{S}_2} \bigcap \mathcal{R}_{\mathcal{S}_3} \bigcap \mathcal{R}_{\mathcal{S}_4} \subseteq \mathcal{R}_{\mathcal{S}_1}$ with respect to $\mathscr{E}$.
- **compat2:** for all $\mathcal{X} \in \mathcal{S}$, $\mathcal{R}_{\mathcal{X}} = \overline{\mathcal{F}}(\mathcal{X})$ with respect to $\mathscr{E}$.

The definition is motivated by the desire to prevent signalling outside of the future. The condition **compat2** identifies the accessible region with the inclusive future, which is based on the ability to broadcast a RV to any location in its future. An alternative would be a weaker condition that requires the accessible region to be some subset of the future. The condition **compat1** is defined in terms of accessible regions, so could also be used with a weaker version of **compat2**. However, a weaker version would in effect place a constraint on broadcasting, and we do not use it here. We return to this in Section V D.

This definition covers all the special cases previously discussed. For single variables, if X affects Y then $\mathcal{Y}$ should be in the future of $\mathcal{X}$ (given **compat2** this is equivalent to taking the accessible region of Y to be contained within that of X); this is **compat1** when $\mathcal{S}_3$ is the empty set (in which case its accessible region is simply $\mathcal{T}$ by Definition V.5) and $\mathcal{S}_1 = \mathcal{X}$ and $\mathcal{S}_2 = \mathcal{Y}$ are single ORVs. When $\mathcal{S}_2$ is a set of ORVs, this case ensures that the ORVs in $\mathcal{S}_2$ are jointly accessible only in the future of the ORV $\mathcal{X}$. This covers the particular case of jamming (Example IV.6).

We now illustrate the definition by applying it to Examples IV.2 to IV.4. In Example IV.2, Z affects Y implies that Y must be in the future of Z and XZ affects Y being a reducible affects relation does not add any further constraints, so we do not require Y to be in the future of X. In Example IV.3, we again must embed Y in the future of Z but in this case, XZ affects Y is irreducible and therefore imposes the constraint that Y must be in the joint future of X and Z. In Example IV.4, in contrast to the previous example, we have X affects Y given $\mathrm{do}(Z)$ even though XZ does not affect Y. The former is irreducible as it involves single RVs, and implies that the joint future of Y and Z must be in the future of X, and since we also have Z affects Y which would require Y to be in the future of Z, we can conclude that compatibility in this case forces Y to be in the future of both X and Z. Noting that W also affects Y, this would require Y to be in the future of W as well.

e. Completeness of Definition V.7: We now provide an argument to show that our definition of compatibility indeed fully captures the intuition of "no signalling outside the future" within our framework. Given **compat2** which we have motivated above, **compat1** is necessary to avoid agents from using the affects relation to signal outside the future, since a violation of compatibility would enable $\mathcal{S}_2$, $\mathcal{S}_3$ and $\mathcal{S}_4$ to be accessed outside the future of $\mathcal{S}_1$ and yet receive a signal from $\mathcal{S}_1$ through the irreducible affects relation $\mathcal{S}_1$ affects $\mathcal{S}_2$ given $\{\mathrm{do}(\mathcal{S}_3), \mathcal{S}_4\}$.[13]

Further if a set of affects relations satisfy our definition with respect to some space-time embedding, this is sufficient to ensure that no agents who can access the associated ORVs can signal outside the future using those affects relations.[14] This is because the conditional HO affects relation $\mathcal{S}_1$ affects $\mathcal{S}_2$ given $\{\mathrm{do}(\mathcal{S}_3), \mathcal{S}_4\}$ (between arbitrary pairwise disjoint sets of ORVs) captures the most general way in which agents can signal to each other in our framework: an agent Alice may intervene on a set $\mathcal{S}_1$ of observed nodes, and an agent Bob with access to another set of observed RVs $\mathcal{S}_2$, can try to detect the effect of Alice's intervention and Bob may additionally have access to some combination of observational ($\mathcal{S}_4$) and interventional data ($\mathrm{do}(\mathcal{S}_3)$) relating to other sets of the observed nodes. Therefore, demanding that $\overline{\mathcal{F}}(\mathcal{S}_2) \bigcap \overline{\mathcal{F}}(\mathcal{S}_3) \bigcap \overline{\mathcal{F}}(\mathcal{S}_4) \subseteq \overline{\mathcal{F}}(\mathcal{S}_1)$ holds for any space-time embedding of the RVs will be sufficient to ensure that this affects relations cannot be used to signal outside the space-time's future. However, this turns out to be too strong a sufficiency condition, and imposing this only for irreducible affects relations (as the definition does) is already sufficient. To see this, suppose that $\mathcal{S}_1$ affects $\mathcal{S}_2$ given $\{\mathrm{do}(\mathcal{S}_3), \mathcal{S}_4\}$ is reducible. Then there exists a subset $s_1 \subset \mathcal{S}_1$ such that s_1 affects $\mathcal{S}_2$ given $\{\mathrm{do}(\mathcal{S}_3), \mathcal{S}_4\}$ (cf. Lemma IV.6). Without loss of generality, take this to be irreducible (if not, simply find a subset of s_1 that satisfies the same affects relation and repeat this argument), then requiring $\overline{\mathcal{F}}(\mathcal{S}_2) \bigcap \overline{\mathcal{F}}(\mathcal{S}_3) \bigcap \overline{\mathcal{F}}(\mathcal{S}_4) \subseteq \overline{\mathcal{F}}(s_1)$ is sufficient to ensure that the reduced affects relation s_1 affects $\mathcal{S}_2$ given $\{\mathrm{do}(\mathcal{S}_3), \mathcal{S}_4\}$ does not signal outside the future. By the reducibility of the original relation $\mathcal{S}_1$ affects $\mathcal{S}_2$ given $\{\mathrm{do}(\mathcal{S}_3), \mathcal{S}_4\}$, we have for $\tilde{s}_1 := \mathcal{S}_1 \backslash \{s_1\}$, $\tilde{s}_1$ does not affect $\mathcal{S}_2$ given $\{\mathrm{do}(\mathcal{S}_3 s_1), \mathcal{S}_4\}$, which means the original affects relation does not require $\overline{\mathcal{F}}(\mathcal{S}_2) \bigcap \overline{\mathcal{F}}(\mathcal{S}_3) \bigcap \overline{\mathcal{F}}(\mathcal{S}_4) \bigcap \overline{\mathcal{F}}(s_1) = \overline{\mathcal{F}}(\mathcal{S}_2) \bigcap \overline{\mathcal{F}}(\mathcal{S}_3) \bigcap \overline{\mathcal{F}}(\mathcal{S}_4)$ to be in contained in the future of $\tilde{s}_1$, once we have imposed $\overline{\mathcal{F}}(\mathcal{S}_2) \bigcap \overline{\mathcal{F}}(\mathcal{S}_3) \bigcap \overline{\mathcal{F}}(\mathcal{S}_4) \subseteq \overline{\mathcal{F}}(s_1)$ for the corresponding reduced relation.

While these arguments justify the completeness of our definition, they do not rule out the possibility of another definition that captures the same intuition. This would also depend on how "no signalling outside the future" is interpreted, and this can be done in several inequivalent ways (e.g., by taking the accessible regions to be a subset of the future in **compat2**), we have proposed one possible, natural way to formalise this. We discuss similar but distinct compatibility conditions in Section V D.

Remark V.2. Given a set $\mathscr{A}$ of arbitrary conditional affects relations (including zeroth and HO relations), one can use the first part of Lemma IV.8 to convert this to a new set $\tilde{\mathscr{A}}$ containing only unconditional affects relations such that compatibility of $\mathscr{A}$ with an embedding $\mathscr{E}$ in a space-time $\mathcal{T}$ implies the compatibility of $\tilde{\mathscr{A}}$ with the same embedding. For this, form $\tilde{\mathscr{A}}$ from $\mathscr{A}$ by including every unconditional affects relation from $\mathscr{A}$, and for every conditional affects relation S_1 affects S_2 given $\{\mathrm{do}(S_3), S_4\}$ in $\mathscr{A}$, add the corresponding unconditional affects relation S_1 affects $\{S_2, S_4\}$ given $\mathrm{do}(S_3)$ in $\tilde{\mathscr{A}}$, if the latter was not already included in $\mathscr{A}$ (note that the former implies the latter by part 1 of Lemma IV.8). Now, every irreducible conditional relation S_1 affects S_2 given $\{\mathrm{do}(S_3), S_4\}$ in $\mathscr{A}$ imposes the condition $\overline{\mathcal{F}}(\mathcal{S}_2) \bigcap \overline{\mathcal{F}}(\mathcal{S}_3) \bigcap \overline{\mathcal{F}}(\mathcal{S}_4) \subseteq \overline{\mathcal{F}}(\mathcal{S}_1)$ on any compatible space-time embedding $\mathscr{E}$. By part 2 of Lemma IV.8, irreducibility of S_1 affects S_2 given $\{\mathrm{do}(S_3), S_4\}$ in $\mathscr{A}$ implies irreducibility of S_1 affects $\{S_2, S_4\}$ given $\mathrm{do}(S_3)$ in $\tilde{\mathscr{A}}$, and the latter imposes the same condition on the embedding, by Definition V.7. Every unconditional relation is present in both sets and hence imply the same conditions on the embedding.

In summary, every affects relation of the form S_1 affects S_2 given $\{\mathrm{do}(S_3), S_4\}$ present in $\mathscr{A}$ can be replaced by S_1 affects $\{S_2, S_4\}$ given $\mathrm{do}(S_3)$ for the purpose of applying Definition V.7.

---

[13] Without **compat2**, **compat1** is necessary for "no signalling outside the accessible region". See Section VII for further discussion.

[14] This applies given the setup assumptions of the framework, such as that interventions are performed independently on each node X and correspond to an exogenous variable I_X etc.

Remark V.3. A complete set of affects relations for a causal model over a set S of RVs is one where for any subsets S_1, S_2, S_3, S_4 of S we know whether or not S_1 affects S_2 given $\{\mathrm{do}(S_3), S_4\}$. It is not always possible to deduce a complete set of affects relations from a causal model (as defined in Definition IV.2), and in general a complete set may not be available. Use of a partial set of affects relations can be sufficient to deduce incompatibility with an embedding, and, given a causal model, a partial set can be deduced. Note that we require causal models to define affects relations in the first place.

Definition V.8 (Compatibility of a causal model with an embedding in a partial order)**.** We say that a causal model over a set of RVs S is compatible with an embedding in a partial order if the set of affects relations $\mathscr{A}$ implied by the causal model are compatible with the embedding (cf. Definition V.7).

Remark V.4. If $X \dashrightarrow Y$, there is no affects relation between X and Y and our compatibility condition does not require that for the corresponding ORV $\mathcal{Y}$, $\mathcal{Y} \in \mathcal{R}_X$. Although demanding this would be natural in light of common notions of causation, one of the motivations behind this line of research is to investigate what happens without this because the existence of such causal influences may not be operationally detectable. In other words, our compatibility condition does not imply that cause precedes effect with respect to the space-time order relation, but it does imply that signalling is not possible outside the future of the space-time structure. Interestingly, this does not rule out the possibility of having causal models with causal loops from being compatibly embedded in the space-time, as we show in Section VI.

D. Necessary and sufficient conditions for compatibility

For compatibility of a set of affects relations with an embedding $\mathscr{E}$ in space-time, Definition V.7 states that the conditions **compat1** and **compat2** must be satisfied. Consider now a similar condition which we call **compat1′**$(\mathcal{S}, \mathscr{A})$, where we use the arguments in the brackets specify the set of ORVs and affects relations that the condition is applied to (since we will later apply it to a different set). With this convention, **compat1**:=**compat1**$(\mathcal{S}, \mathscr{A})$.

compat1′$(\mathcal{S}, \mathscr{A})$**:** Let $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathcal{S}$ be disjoint proper subsets of ORVs, and $\mathcal{S}_3, \mathcal{S}_4$ be two other subsets (possible empty) disjoint from themselves and $\mathcal{S}_1$ and $\mathcal{S}_2$. If $\mathcal{S}_1$ affects $\mathcal{S}_2$ given $\{\mathrm{do}(\mathcal{S}_3), \mathcal{S}_4\}$ is in $\mathscr{A}$ and is irreducible with respect to the affects relations in $\mathscr{A}$, then $\bigcap_{s_{234} \in \mathcal{S}_2\mathcal{S}_3\mathcal{S}_4} \overline{\mathcal{F}}(s_{234}) \subseteq \bigcap_{s_1 \in \mathcal{S}_1} \overline{\mathcal{F}}(s_1)$ with respect to $\mathscr{E}$.

Note that **compat1′**$(\mathcal{S}, \mathscr{A})$ imposes no condition on the accessible regions but only on the space-time locations of the ORVs (which allow us to fully specify their inclusive futures), while **compat1** restricts the accessible regions. However, once **compat2** is imposed, **compat1** and **compat1′**$(\mathcal{S}, \mathscr{A})$ are essentially equivalent i.e., an equivalent definition of compatibility would be to use **compat1′**$(\mathcal{S}, \mathscr{A})$ and **compat2** instead of **compat1** and **compat2** in Definition V.7. We use **compat1** instead of **compat1′**$(\mathcal{S}, \mathscr{A})$ in the original definition to make it clear that this condition is related to the operational concept of "accessibility" of ORVs, which is captured by the accessible regions. In general, the accessible region of an ORV need not be fully specified by its space-time location or even be related to its future, but this is the case once **compat2** is assumed. The following theorem (proven in Appendix D 3) and corollary establish certain useful connections between these concepts, and follow from Definition V.7.

Theorem V.1. *[Necessary and sufficient conditions for compatibility with an embedding in $\mathcal{T}$] Let $\mathcal{S}$ be set of ORVs embedded in a partial order $\mathcal{T}$ with respect to an embedding $\mathscr{E}$ and let $\mathscr{A}$ be a given set of affects relations on $\mathcal{S}$. Further, consider forming an augmented set of ORVs $\mathcal{S}'$ by taking $\mathcal{S}$ and for each variable $\mathcal{X} \in \mathcal{S}$, embedding a copy of $\mathcal{X}$ at each point in its accessible region $\mathcal{R}_\mathcal{X}$ and form $\mathscr{A}'$ by adding to $\mathscr{A}$ that each variable affects each of its copies for all copies. Then the following statements hold.*

1. *If the set of affects relations $\mathscr{A}$ is compatible with the embedding $\mathscr{E}$ in $\mathcal{T}$, then* **compat1′**$(\mathcal{S}', \mathscr{A}')$ *holds i.e.,* **compat1′**$(\mathcal{S}', \mathscr{A}')$ *is necessary for compatibility of $\mathscr{A}$ with the space-time embedding $\mathscr{E}$.*
2. **compat1′**$(\mathcal{S}', \mathscr{A}')$ *implies that $\mathcal{R}_\mathcal{X} \subseteq \overline{\mathcal{F}}(\mathcal{X})$ $\forall \mathcal{X} \in \mathcal{S}$, but not that the two sets $\mathcal{R}_\mathcal{X}$ and $\overline{\mathcal{F}}(\mathcal{X})$ are necessarily equal $\forall \mathcal{X} \in \mathcal{S}$, i.e.,* **compat1′**$(\mathcal{S}', \mathscr{A}')$ *is not sufficient for compatibility of $\mathscr{A}$ with the space-time embedding $\mathscr{E}$.*

The augmented sets $\mathcal{S}'$ and $\mathscr{A}'$ in the above theorem capture the idea of broadcasting classical RVs to each point in their accessible region. Imposing **compat1′** for the an embedding $\mathscr{E}$ of these sets in space-time then ensures that this broadcasting (i.e., finding copies of the RVs) is possible only within the future, but not necessarily to all locations in the future. Note that being able to find copies of an ORV $\mathcal{X}$ only within its future does not by itself imply that any ORV $\mathcal{Y}$ affected by X must be contained in its future. We then have the following corollary of the theorem.

Corollary V.1. *Let S be a set of RVs and $\mathscr{A}$ be a set of affects relations over them. Then there exists a non-trivial embedding $\mathscr{E}$ of S in a partial order $\mathcal{T}$ compatible with $\mathscr{A}$ if and only if there exists a non-trivial embedding $\mathscr{E}'$ of the same affects relations that satisfies* **compat1′**$(\mathcal{S}', \mathscr{A}')$.

That the existence of a non-trivial embedding $\mathscr{E}$ that satisfies **compat** implies the existence of one that satisfies **compat1′**$(\mathcal{S}', \mathscr{A}')$ follows directly from the necessary part of Theorem V.1. The other direction follows because any non-trivial embedding $\mathscr{E}'$ that satisfies **compat1′**$(\mathcal{S}', \mathscr{A}')$ can be turned into a non-trivial embedding $\mathscr{E}$ that satisfies **compat** simply by taking $\mathscr{E}'$ and setting the accessible regions of ORVs to satisfy **compat2**. The important point to note is that the two embeddings $\mathscr{E}$ and $\mathscr{E}'$ need not be the same.

VI. CAUSAL LOOPS AND THEIR SPACE-TIME EMBEDDINGS

We have characterised a general class of causal models, defined when a given causal model can be said to be compatible with a space-time embedding and also compared related yet distinct conditions on the space-time embeddings. It is interesting to consider whether there are certain structural properties of the causal model alone that guarantee the existence of a non-trivial and compatible space-time embedding for that causal model. Clearly, the acyclicity of the causal structure is such a property, while this is certainly sufficient, a natural question is whether is it also necessary to guarantee the existence of such a space-time embedding. This question motivates us to define a broad set of possible theories $\mathbb{T}$ that are consistent with the principle of "no signalling outside the future". The set $\mathbb{T}$ consists of theories with the property that for every causal models that can arise in the theory, there exists a non-trivial and compatible embedding in a space-time (cf. Definition V.7).

This class of theories is quite general, it certainly includes quantum and standard GPTs and any theory that can be characterised using acyclic causal structure. In the associated Letter [1], we apply the framework developed here to construct an explicit operationally detectable causal loop that can be embedded in (1+1)-dimensional Minkowski space-time without superluminal signalling, which demonstrates that the set $\mathbb{T}$ can also include theories admitting causal loops. In this section, we characterise several different classes of causal loops that can arise in our framework, and we show that some of these classes can be ruled out by requiring that the causal model has a compatible space-time embedding while the results of the associated Letter show that some other classes cannot be ruled out in this manner. We provide further examples to argue that fully characterising the set of theories $\mathbb{T}$ may be a difficult task. By full characterisation, we mean finding a necessary and sufficient set of conditions on the set of possible affects relations (and/or correlations) of the causal model that guarantees the existence of a non-trivial compatible space-time embedding. Let us now take a closer look at the types of causal loops that can arise in our framework.

A. Different classes of causal loops

We have seen that due to fine-tuning, causation does not imply the existence of affects relations. This motivated the classification of causal arrows (Definition IV.4) into solid and dashed based on the existence of suitable affects relations. Similarly, we can distinguish between different types of causal loops in our framework depending on whether they can be operationally detected through their affects relations. A causal loop simply corresponds to a directed cycle in a causal structure $\mathcal{G}$ involving at least two observed nodes i.e., two observed nodes X and Y in $\mathcal{G}$ such that there exist directed paths from X to Y and from Y to X. Often however, we may not know the full causal structure but only a set of affects relations $\mathscr{A}$ over the observed nodes of an underlying causal structure $\mathcal{G}$. The set $\mathscr{A}$ might allow us to infer some, but not necessarily all the causal relationships in $\mathcal{G}$. We then have the following two broad categories of causal loops, the former (affects causal loops) are operationally detectable via the their affects relations and the latter (hidden causal loops) are not operationally detectable through their affects relations or correlations.

Definition VI.1 (Affects causal loops (ACL))**.** Any set of affects relations $\mathscr{A}$ that can only arise in a causal model associated with a cyclic causal structure $\mathcal{G}$ are said to form/contain an affects causal loop. In other words, affects causal loops certify the cyclicity of the underlying causal structure through the observed affects relations.

Definition VI.2 (Hidden causal loop (HCL))**.** Given a causal model whose causal structure contains a directed cycle, and a complete set of affects relations, we say that this causal model contains a *hidden causal loop* if the same set of affects relations and the same correlations are also realisable in an acyclic causal structure.

A HCL is by definition a causal loop since it corresponds to a directed cycle in the causal structure. These act as causal loops at the level of the causal mechanisms but cannot be detected at the operational level of affects relations (or correlations). It can be the case that causal structures contain directed cycles without being an ACL, meaning

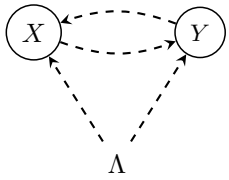


FIG. 11: An operationally undetectable causal loop (Example VI.1).

that the affects relations of the associated causal model can also be obtained in an acyclic causal structure. This does not necessarily imply that both the affects relations and correlations can be generated in an acyclic causal structure, so ACLs and HCLs are not complements of one another. Below we provide an example of a HCL.

Example VI.1 (An operationally undetectable causal loop)**.** Consider the causal structure of Figure 11 over the binary RVs X, Y and Λ, where X and Y are observed nodes which are causes of each other (forming a causal loop) and Λ is an unobserved common cause of the two. Suppose that the RVs are related as follows: Λ is uniformly distributed and $X = \Lambda \oplus Y$ and $Y = \Lambda \oplus X$. Note that the given causal structure already implies a complete set of affects relations i.e., for each pair of the observed nodes, we know whether or not one affects the other. In this case, this is implied by the dashed arrows and we have that X and Y do not affect each other. Since Λ is uniform, $P_{\mathcal{G}_{\mathrm{do}(X)}}(Y|X)$ and $P_{\mathcal{G}_{\mathrm{do}(Y)}}(X|Y)$ are both uniform, and in order to have the required (non-)affects relations, it must be that case that $P_{\mathcal{G}}(X)$ and $P_{\mathcal{G}}(Y)$ are both uniform. Along with the given functional dependences, this implies that X and Y are uncorrelated with each other. In other words, there are no affects relations or correlations between the set of observed nodes of this causal structure even though there is a causal loop. A causal structure over X and Y with no edges at all would also explain these observations. Therefore the directed cycle between X and Y in Figure 11 corresponds to a hidden causal loop. It is also worth noting that knowing the value of the exogeneous variable Λ is not enough to determine the value of X or Y with the given functional relations; in this sense the causal model appears incomplete.

We now focus on the more interesting class of causal loops, affects causal loops. Definition VI.1 only tells us that these are causal loops whose existence is operationally certified by the observable affects relations. It is natural to seek necessary and sufficient conditions on the set of affects relations such that they form an ACL. Here (and in Appendix B), we propose several sufficient conditions which can be considered as definitions of different types of affects causal loops. We discuss 6 types here and 4 more in the appendix and provide examples to illustrate that none of these are necessary conditions i.e., there can be further types of ACLs not covered by these ten types. After defining the 6 types here, we will prove that these are indeed ACLs (Theorem VI.1).

A first sufficient condition for the existence of an ACL is that there are two RVs X and Y that affect each other. Since affects implies cause, this tells us that X and Y must be causes of each other and hence that these affects relations are only realisable in a cyclic causal structure i.e., they lead to an ACL. A second condition is the presence of a chain of single RV affects relations from X to Y and from Y to X. The latter can in general be a distinct condition from the former due to the non-transitivity of the affects relation (see Example IV.1), but can be shown to be an ACL. This gives us the following two types of causal loops.

Definition VI.3 (Affects causal loops, Type 1 (ACL1))**.** A set of affects relations $\mathscr{A}$ is said to contain a Type 1 affects causal loops if there exist two RVs X and Y such that $\{X \text{ affects } Y, Y \text{ affects } X\} \subseteq \mathscr{A}$.

Definition VI.4 (Affects causal loops, Type 2 (ACL2))**.** A set of affects relations $\mathscr{A}$ is said to contain a Type 2 affects causal loop if there exist RVs X, Z_1, Z_2, ..., Z_k and Y such that X affects Z_1, Z_1 affects Z_2, ..., Z_k affects Y and Y affects X are all in $\mathscr{A}$.

More generally, one can also consider affects relations involving sets of RVs. A first observation is that S_1 affects S_2 and S_2 affects S_1 for two sets of RVs does not imply the existence of a directed cycle in the causal structure. For example, consider a causal structure $\mathcal{G}$ with 4 nodes A, B, C and D, all of which are observed such that the only edges in $\mathcal{G}$ are the solid arrows $A \longrightarrow B$ and $C \longrightarrow D$, with A affects B and C affects D. Then, if $S_1 = AD$ and $S_2 = BC$ we have S_1 affects S_2 and S_2 affects S_1 even though $\mathcal{G}$ is clearly acyclic. However, if we take these to be irreducible affects relations, this will no longer be the case and we can certify the cyclicity of the causal structure from the affects relations, as we later show. This motivates more general set of sufficient conditions for the existence of affects causal loops. Two immediate possibilities are the following.

Definition VI.5 (Affects causal loops, Type 3 (ACL3))**.** A set of affects relations $\mathscr{A}$ is said to contain a Type 3 affects causal loop if there exist two disjoint sets S_1 and S_2 of RVs such that $\{S_1 \text{ affects } e_2, S_2 \text{ affects } e_1\} \subseteq \mathscr{A}$ where $e_1 \in S_1$, $e_2 \in S_2$, and both affects relations are irreducible.

Definition VI.6 (Affects causal loops, Type 4 (ACL4))**.** A set of affects relations $\mathscr{A}$ is said to contain a Type 4 affects causal loop if there exist sets of RVs S_1, S_2, …, S_n where each pair S_i and $S_{i+1 \mathrm{mod} n}$ is disjoint, such that $\{S_1$ affects S_2, S_2 affects S_3, …, S_{n-1} affects S_n, S_n affects $S_1\} \subseteq \mathscr{A}$, and all these affects relations are irreducible.

ACL1, ACL2, ACL3 and ACL4 imply cyclicity of the causal structure as shown in Theorem VI.1. However, these are not the most general conditions on the affects relations with this property. There can be further conditions that are not equivalent to ACL1, ACL2, ACL3 or ACL4 which also imply cyclicity. The following is such a condition.

Definition VI.7 (Affects causal loops, Type 5 (ACL5))**.** A set of affects relations $\mathscr{A}$ is said to contain a Type 5 affects causal loop if there exist sets of RVs $S_i \subseteq \hat{S}_i$ for $i = 1, \ldots, n$ such that $\hat{S}_1$ affects S_2, $\hat{S}_2$ affects S_3, …, $\hat{S}_{n-1}$ affects S_n and $\hat{S}_n$ affects S_1 are all in $\mathscr{A}$, where all the affects relations are irreducible and every pair of sets connected by an affects relation is disjoint. Such a chain of affects relations is called a complete affects chain, in this case the affects chain is from the set S_1 to itself.

Rather than considering a chain of irreducible affects relations from an RV or a set of RVs back into itself, one can consider multiple chains which taken together imply cyclicity and this would give yet another type of causal loop in our framework. For example we may have an irreducible affects relation A affects BC. Along with another irreducible affects relation BCD affects A, this would form a Type 5 affects causal loop. By Corollary IV.3, these affects relations would tell us that A is either a cause of B or C while B, C and D are all causes of A. Irrespective of whether A is a cause of B or of C, this implies the existence of a directed cycle in the causal structure. However, we could instead have started with the irreducible affects relations A affects BC, B affects A and C affects A. Since in general, B affects A, and C affects A need not imply BC affects A (see Example IV.4), these affects relations may not constitute a Type 5 affects loop but they nevertheless imply cyclicity (using Corollary IV.3). Note that A affects BC and B affects A alone (even if irreducible) do not necessarily imply cyclicity since the former tells us that A is either a cause of B or of C and the latter that B is a cause of A. That is, these affects relations can in principle be obtained in an acyclic causal model where A is a cause of C and B is a cause of A. Generalising this idea, we have another type of affects loop, ACL6.

Definition VI.8 (Affects causal loops, Type 6 (ACL6))**.** A set of affects relations $\mathscr{A}$ is said to contain a Type 6 affects causal loop if the following conditions are satisfied

1. There exist disjoint sets of RVs S_1 and S_2 such that S_1 affects S_2 belongs to $\mathscr{A}$ and is irreducible.
2. For each element $e_2 \in S_2$, there exists a *complete chain of irreducible affects relations* that connects it back to S_1, i.e., for each e_2, there exists sets of RVs $S_i \subseteq \hat{S}_i$ for $i = 1, \ldots, n$ and $s_1 \subseteq S_1$ such that $\{\hat{S}_2$ affects S_3, $\hat{S}_3$ affects S_4, …, $\hat{S}_{n-1}$ affects S_n, $\hat{S}_n$ affects $s_1\} \subseteq \mathscr{A}$, where all the affects relations are irreducible and every pair of sets connected by an affects relation is disjoint.

There are further types of affects causal loops, all of which imply cyclicity of the causal structure. For example, we can also consider affects causal loops involving chains of conditional higher-order affects relations (Definition IV.5) and define analogues of ACL1-6 for this case. These can in general be distinct from ACL1-6 since it is possible to have a conditional HO affects relation X affects Y given $\{\mathrm{do}(Z), W\}$ without the unconditional zeroth-order affects relation X affects Y. Even using unconditional zeroth-order affects relations alone, further distinct classes of affects causal loops are possible and four such classes (ACL7 to ACL10) are described in Appendix B. The intuition behind them is as follows. The kind of chains of irreducible affects relations considered in the above definitions are such that for each subsequent pair of affects relations $\hat{S}_i$ affects S_{i+1}, the set S_{i+1} is contained in $\hat{S}_{i+1}$. What if this were not the case, and we only had that $S_{i+1} \bigcap \hat{S}_{i+1} \neq \varnothing$? Let us call this an "incomplete" affects chain. The example before the last definition, with A affects BC and B affects A illustrates that this condition alone is not enough to guarantee cyclicity and to justify calling these affects relations a causal loop. One way is to add the affects relation C affects A, which motivates the definition of ACL6 above. Another option is to add the irreducible affects relations C affects D and D affects BC and one can again show that the set of irreducible relations $\mathscr{A} = \{A$ affects BC, B affects A, C affects D, D affects $BC\}$ is cyclic. One can however verify that this $\mathscr{A}$ does not correspond to any of the affects causal loops previously defined. There are two incomplete affects chains that complete each other, but no complete chain as required by the above types of ACL. In general, one might need to combine a given incomplete chain with several other complete or incomplete chains to guarantee cyclicity of the resulting set of affects relations, and the conditions therefore continue to get more complex. Even the additional classes of affects causal loops defined in Appendix B do not exhaust all the possible types of affects causal loops that might be possible in our framework (we provide an example in the appendix to illustrate this).

The following theorem (proven in Appendix D 3) shows that ACL1-6 are indeed affects causal loops in the sense of Definition VI.1.

Theorem VI.1. *Any set of affects relations $\mathscr{A}$ containing an affects causal loop of Type 1, 2, 3, 4, 5 or 6 can only arise from a causal model over a cyclic causal structure i.e., these are indeed instances of affects causal loops according to Definition VI.1.*

B. Possibility of compatibly embedding causal loops in space-time

In the previous section we discussed various properties of causal loops that follow from the causal model alone and without reference to space-time. Here we consider the space-time embeddings of such loops and whether affects causal loops can be compatibly and non-trivially embedded in a space-time structure. This turns out to indeed be possible for certain types of affects causal loops. This implies that for some causal loops their existence can be operationally certified (through observed affects relations, by virtue of being affects causal loops), and they can nevertheless be non-trivially embedded in space-time without leading to signalling outside the space-time future. While our framework can be applied to arbitrary partially ordered space-times, for the sake of illustration, we consider the case of (1+1)-dimensional Minkowski space-time in this section. Before we show the existence of embeddable causal loops in this case, we make the following observation.

Lemma VI.1. *Let S be a set of RVs and $\mathscr{A}$ be a set of affects relations over them.*

1. *The absence of affects causal loops (Definition VI.1) in $\mathscr{A}$ is a sufficient condition for the existence of a non-trivial embedding of S in a space-time that $\mathscr{A}$ is compatible with.*

2. *If $\mathscr{A}$ is assumed to be a set of affects relations associated with a faithful causal model, then all causal loops are Type 1 affects causal loops and the existence of a non-trivial space-time embedding of S that $\mathscr{A}$ is compatible with is both necessary and sufficient to rule out all causal loops and guarantee the acyclicity of the causal model that generates $\mathscr{A}$.*

The above lemma (proven in Appendix D 3) shows that all the distinct classes of causal loops ACL2 to ACL6 (and ACL7 to ACL10 and other possible classes as described in Appendix B) as well as the concept of hidden causal loops only arise in fine-tuned causal models. If fine-tuning is allowed, even the absence of affects causal loops does not rule out causal loops since we can have hidden causal loops which are operationally undetectable i.e., the absence of ACL does not imply acyclicity of the causal structure. Here, we first show that the absence of Type 1 and Type 2 affects causal loops is necessary for the existence of such a non-trivial and compatible space-time embedding. The results of the associated Letter [1] show that this is no longer true for ACLs of higher types, in particular we construct an ACL of Type 4 there that does admit such a space-time embedding. This demonstrates that the absence of affects causal loops is not necessary for the existence of a non-trivial and compatible space-time embedding. We further show here that the absence of Type 1 and 2 loops is not sufficient for the existence of a non-trivial and compatible space-time embedding, since such an embedding is not guaranteed to exist for affects loops of other types i.e., for ACL3 and above there may or may not exist a non-trivial and compatible space-time embedding (this is discussed in Appendix B).

Consider the affects causal loops of Types 1 and 2. Recall that a non-trivial space-time embedding is one where no two RVs such that one affects the other are assigned the exact same space-time location. A non-trivial space-time embedding is impossible for ACL1 and ACL2, since **compat** applied to a set of affects relations containing an ACL2 implies that $\mathcal{X} \preceq \mathcal{Z}_1 \preceq \ldots \preceq \mathcal{Z}_k \preceq Y \preceq \mathcal{X}$ which can only be satisfied when $\mathcal{X} \preceq \mathcal{Y} \preceq \mathcal{X}$ i.e., $O(X) = O(Y)$, which corresponds to a trivial embedding. The latter step follows directly by applying **compat** for ACL1. This is stated explicitly in the following Lemma.

Lemma VI.2. *Let S be a set of RVs and $\mathscr{A}$ be a set of affects relations over them that contains affects causal loops of Types 1 or 2. The set S cannot be non-trivially embedded in any space-time such that $\mathscr{A}$ is compatible with the embedding.*

Now consider ACL3 formed by the irreducible affects relations $\mathscr{A} = \{AB \text{ affects } C,\ CD \text{ affects } A\}$. Applying **compat** to the first affects relation, we have that C must be in the joint inclusive future of A and B i.e., $\mathcal{A} \preceq \mathcal{C}$ and $\mathcal{B} \preceq C$. The condition **compat** for the second affects relation similarly implies that $\mathcal{C} \preceq \mathcal{A}$ and $\mathcal{D} \preceq \mathcal{A}$. Together these imply that A and C must be embedded at the same location while B and D cannot be in the future of this location. Since we neither have A affects C nor C affects A in $\mathscr{A}$, there is a non-trivial embedding. However, if we form $\mathscr{A}'$ by adding one or both of these affects relations to $\mathscr{A}$, there will no longer be any non-trivial and compatible embedding. In other words, affects causal loops of Type 3 can admit non-trivial and compatible space-time embeddings, but will always be *degenerate*, i.e., they require two of the RVs to be embedded at the same location (e_1 and e_2 in Definition VI.5), as shown in the lemma below.

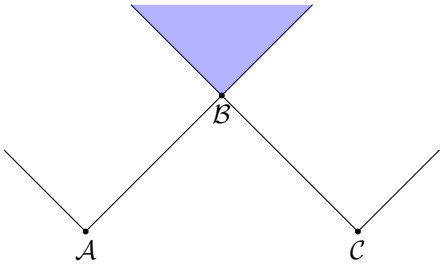


FIG. 12: **A non-trivial and compatible space-time embedding for an operationally detectable causal loop** Example VI.2 describes a set of affects relations that forms an affects causal loop of Type 4. Such a causal loop is operationally detectable since the cyclicity of the underlying causal model can be certified operationally using the observed affects relations, as shown in Theorem VI.1. This figure illustrates a non-trivial and non-degenerate, yet compatible embedding of this causal loop in (1+1)-dimensional Minkowski space-time, where space and time are given along the horizontal and vertical axes respectively and black lines correspond to light cones. Note that this embedding remains compatible even when the space-time RVs $\mathcal{A}$ and $\mathcal{C}$ are pushed to the far past of $\mathcal{B}$ along the black line ($\mathcal{B}$'s past light-like surface).

Lemma VI.3. *Let S be a set of RVs and $\mathscr{A}$ be a set of affects relations over them that contains affects causal loops of Type 3. The set S cannot be embedded in any space-time such that the embedding is non-degenerate such that $\mathscr{A}$ is compatible with the embedding. However, there are non-trivial embeddings that $\mathscr{A}$ is compatible with.*

Proof. By Definition VI.5, ACL3 implies that for two sets S_1 and S_2 of RVs, we have the irreducible affects relations S_1 affects e_2 and S_2 affects e_1 for some elements $e_1 \in S_1$ and $e_2 \in S_2$. Applying **compat** (Definition V.7), this implies that e_1 must be embedded in the inclusive future of all elements $e_2' \in S_2$ and e_2 must be embedded in the inclusive future of all elements $e_1' \in S_1$. This is only possible if e_1 and e_2 are embedded at the same location, making the embedding degenerate. However, it can be the case that $\mathscr{A}$ does not contain or imply the any affects relations between e_1 and e_2, therefore the embedding may still be non-trivial. □

Can we embed an affects causal loop compatibly in space-time such that all RVs have distinct locations? The associated Letter [1] shows that such a non-degenerate embedding is indeed possible for certain types of affects causal loops, with an explicit example. The causal loop proposed in [1] corresponds to a Type 4 ACL in the language of the present paper, we reproduce this example here completeness.

Example VI.2 (An operationally detectable causal loop with a non-trivial, compatible space-time embedding [1])**.** Suppose we have the irreducible affects relations $\mathscr{A} = \{B \text{ affects } AC,\ AC \text{ affects } B\}$ which form a Type 4 ACL. Then **compat** would require that $\overline{\mathcal{F}}(\mathcal{B}) = \overline{\mathcal{F}}(\mathcal{A}) \bigcap \overline{\mathcal{F}}(\mathcal{C})$. This can be satisfied even when A, B and C are embedded at distinct space-time locations, as shown in Figure 12. This figure shows that this affects causal loop involving the RVs A, B and C can be embedded in a (1+1)-dimensional Minkowski space-time without leading to signalling outside the space-time future. This is possible even if we embed A and C arbitrarily far in the past, as long as the earliest location where their lightcones intersect coincides with the location of B. By Theorem VI.1, observation of the affects relations $\{B \text{ affects } AC,\ AC \text{ affects } B\}$ operationally certifies the existence of a causal loop i.e., that there exist at least one pair of RVs among A, B and C that are causes of each other. This causal loop corresponds to a closed timelike curve (CTC) once the RVs are embedded in a space-time, since it would imply bidirectional causal influences between two distinct space-time locations. Even if this CTC involves causal influences between RVs that occur far apart in time (in some reference frame), they will not allow any agent to signal superluminally since the affects relations are compatible with the space-time. This is true even if the agent can access all the RVs or any subset thereof. This is because both of the affects relations in $\mathscr{A}$ can only be verified only in the joint future of A and C, and the earliest point that they can do so is the location of B.

An explicit cyclic causal model in which the affects causal loop of the above example can arise is also provided in [1], we further discuss this in Appendix A and Figure 14b along with other examples that illustrate the concept of "higher order affects relations" introduced in this work. One can also use the framework developed here to construct several further examples of causal loops (of ACL4 or higher types) that can be compatibly embedded in space-time. The example provided in our Letter [1] suffices to illustrate the claim that such loops are even possible. We discuss further, the space-time embeddings of higher types of ACLs in Appendix B.

Remark VI.1 (Types of space-time embeddings)**.** Apart from distinguishing between different types of causal loops (that arise due to fine-tuning of the underlying parameters of the causal model), one might also wish to distinguish

between different types of space-time embeddings. Some useful distinctions that were made so far are between trivial and non-tivial embeddings and degenerate and non-degenerate embeddings. The former is useful because any set of affects relations can be compatibly embedded in a space-time through a trivial embedding where all RVs are embedded at the same location, and this does not tell us anything interesting. If we demand non-trivial embeddings, i.e., that two RVs connected by an affects relation only involving them are not embedded at the same location, then this rules out affects causal loops of Types 1 and 2, as shown in Lemma VI.2 but not Type 3 loops. On the other hand, if we demand non-degenerate embeddings, i.e., that all RVs are embedded at distinct locations, we can rule out Type 3 affects causal loops as shown in Lemma VI.3, but not Type 4. Note that the compatible and non-degenerate space-time embedding of the Type 4 affects causal loop that we propose in the Letter [1] (and discussed in Example VI.2) is "fine-tuned" and is *unstable* in the sense that small adjustments to the space-time embedding of the variables would break compatibility. In the case of Minkowski space-time, the requirement $\overline{\mathcal{F}}(\mathcal{B}) = \overline{\mathcal{F}}(\mathcal{A}) \cap \overline{\mathcal{F}}(\mathcal{C})$ that guarantees compatibility of the ACL4 in Example VI.2 confines the ORVs $\mathcal{A}$ and $\mathcal{C}$ to a surface that is one dimension smaller than the dimensions of the space-time, once the location of the ORV $\mathcal{B}$ is fixed. This surface if simply the boundary of the past light cone of $\mathcal{B}$. $\mathcal{A}$ and $\mathcal{C}$ can be placed anywhere on this surface, including arbitrarily far in the past of $\mathcal{B}$ along its past light-like surface but cannot be placed out of this surface without violating compatibility. [Alternatively, once $\mathcal{A}$ and $\mathcal{C}$ are embedded, there is only one possible location for $\mathcal{B}$.] Such examples of causal loops that do not lead to superluminal signalling involve a form of fine-tuning both at the level of the causal model and at the level of its space-time embedding.

Remark VI.2 (Open questions and challenges)**.** As motivated in the above remark, one can consider further distinctions between space-time embeddings, such as whether they are unstable embeddings. We have seen that such embeddings arise in Example VI.2 and other examples of this section and Appendix B. All the non-degenerate and compatible space-time embeddings of affects causal loops that we know so far (such as Example VI.2) are such unstable embeddings. Therefore an interesting open question is whether demanding that an embedding is *stable* would rule out some or all of the affects causal loops of Types 4 and higher.

It remains unclear what condition on the space-time embedding would rule out all possible types of affects causal loops. A main reason is that the general class of affects causal loops (i.e., operationally detectable causal loops) is not fully characterised, ACL1–ACL6 only provide various sufficient conditions that imply the existence of an affects causal loop but none of them, including the further classes ACL7–ACL10 discussed in Appendix B are necessary. In all the classes other than ACL1 and ACL2, one can find causal loops that admit non-trivial and compatible space-time embeddings, but it is also possible to find ACLs of other types that have no non-trivial or compatible embeddings. Thus the question regarding necessary and sufficient conditions on affects relations that guarantee a non-trivial and compatible embedding (and similarly for other types of embeddings) also remains open.

While we have seen that there is a non-trivial and compatible embedding of the affects loop of Example VI.2 in (1+1)-dimensional Minkowski space-time, there is no such embedding of the same loop in (3+1)-dimensional Minkowski space-time [1]. This is because the compatibility condition requires B to be embedded at the earliest location in the joint future of A and C which is not possible in (3+1)-dimensional Minkowski space-time where a frame-dependent concept of earliest location in the joint future does not exist (in contrast to the (1+1)-dimensional case). This implies that the conditions for ruling out causal loops in a space-time can depend on the dimension of the space-time, and possibly other topological features. In particular, it remains a pertinent open question whether the existence of a non-trivial and compatible embedding in the space-time is sufficient to rule out all affects causal loops in (3+1)-dimensional Minkowski space-time. We leave these open questions as a challenge for future research in the field.

The framework developed here, along with the results of the associated Letter [1] illustrate the counter-intuitive possibilities offered by fine-tuning—if it is possible to have superluminal causal influences without superluminal signalling (as in non-local hidden variable theories [37] or the jamming scenario [30, 31]), then we can also have causal loops that do not lead to superluminal signalling. The particularly interesting feature of such causal loops is that they can be operationally detected through their affects relations. These results have consequences for the claims of [30, 31] that certain constraints on correlations in Bell scenarios are necessary and sufficient for ruling out all types of causal loops. They suggest that neither directions of these claims can hold. This is discussed in the following section.

VII. CRITICAL ANALYSIS OF PREVIOUS CLAIMS REGARDING RELATIVISTIC CAUSALITY

Here we comment on two related works, [30] where the concept of jamming non-local correlations were introduced and [31] where these were further analysed and generalised to so-called "relativistic causal correlations". The results and assumptions of [30, 31] are not stated in the same mathematical language as ours, and hence some translation is needed to use our framework. For this discussion, we consider a tripartite Bell experiment, i.e., we consider six

random variables: the settings (A, B, C), and corresponding outcomes (X, Y, Z) that are embedded into Minkowski space-time satisfying the following constraints.

Definition VII.1 (Embeddings of the form $\mathscr{E}^{\text{jam}}$)**.** Random variables A, B, C, X, Y and Z have an embedding of the form $\mathscr{E}^{\text{jam}}$ if the following conditions are satisfied

1. $\{\mathcal{A},\mathcal{X}\} \not\asymp \{\mathcal{B},\mathcal{Y}\}$, $\{\mathcal{A},\mathcal{X}\} \not\asymp \{\mathcal{C},\mathcal{Z}\}$ $\{\mathcal{B},\mathcal{Y}\} \not\asymp \{\mathcal{C},\mathcal{Z}\}$, $\mathcal{A} \preceq \mathcal{X}$, $\mathcal{B} \preceq \mathcal{Y}$, $\mathcal{C} \preceq \mathcal{Z}$,[15]
2. $\overline{\mathcal{F}}(\mathcal{X}) \cap \overline{\mathcal{F}}(\mathcal{Z}) \subseteq \overline{\mathcal{F}}(\mathcal{B})$

The first of these conditions corresponds to space-time constraints for a tripartite Bell scenario. A, B and C can be thought of as settings with X, Y and Z corresponding outcomes. The first condition then represents the space-like separation of the three parts of the experiment, with each setting embedded in the space-time past of the corresponding outcome. The second condition is an additional restriction on the space-time location of the RVs related to the particular jamming scenario we wish to consider, demanding that the joint future of X and Z is in the future of B. These conditions define a family of embeddings; the locations of A, B, C, X, Y and Z in Figure 3 satisfy the conditions. Tripartite Bell experiments carried out in space-like separated configurations satisfying the first conditions of $\mathscr{E}^{\text{jam}}$ are normally associated with a set of no-signalling constraints on the possible correlations. However, given both conditions of $\mathscr{E}^{\text{jam}}$, the works [30, 31] consider a relaxed set of no-signalling conditions, as follows.

Definition VII.2 (Relaxed tripartite no-signalling conditions [30] (**NS3′**))**.** The relaxed no-signalling conditions **NS3′** associated with an embedding of the form $\mathscr{E}^{\text{jam}}$ correspond to the following constraints on the observed distribution $P_{XYZ|ABC}$.

$$\begin{aligned}
P_{XY|AB}(x,y|a,b) &:= \sum_z P_{XYZ|ABC}(x,y,z|a,b,c) = \sum_z P_{XYZ|ABC}(x,y,z|a,b,c') \quad \forall x,y,a,b,c,c' \\
P_{YZ|BC}(y,z|b,c) &:= \sum_x P_{XYZ|ABC}(x,y,z|a,b,c) = \sum_x P_{XYZ|ABC}(x,y,z|a',b,c) \quad \forall y,z,a,a',b,c \\
P_{X|A}(x|a) &:= \sum_{y,z} P_{XYZ|ABC}(x,y,z|a,b,c) = \sum_{y,z} P_{XYZ|ABC}(x,y,z|a,b',c') \quad \forall x,a,b,b',c,c' \\
P_{Z|C}(z|c) &:= \sum_{x,y} P_{XYZ|ABC}(x,y,z|a,b,c) = \sum_{x,y} P_{XYZ|ABC}(x,y,z|a',b',c) \quad \forall z,a,a',b,b',c
\end{aligned} \tag{9}$$

Note that these conditions imply $P_{Y|ABC}(y|abc)$ is independent of a and c, so that $P_{Y|B}$ is well defined. The idea behind these relaxed conditions is that they allow $P_{XZ|ABC}$ to depend on B (which would normally be forbidden) on the grounds that the joint future of X and Z is contained in that of B in the embedding $\mathscr{E}^{\text{jam}}$, and hence information about B can remain in its future (as explained in Section III B).

Since the conditions **NS3′** involve only the observed correlations, they do not by themselves tell us about causation. Therefore, without making further assumptions about the underlying causal model, they cannot be necessary and sufficient conditions to rule out superluminal signalling or causal loops. For instance, a set of correlations violating **NS3′** could arise from a single unobserved common cause of all six variables without any direct causes, which would not lead to any superluminal signalling. When referring to such conditions on correlations as "no-signalling" conditions, we often implicitly assume some notion of "free choice" for the settings (see [31, 47, 48] for definitions of free choice). In the causal modelling framework, free choice can be modelled by taking the settings A, B and C to be exogenous i.e., as having no prior causes. Given the exogeneity of A, B and C, **NS3′** capture the signalling possibilities through interventions on these variables (cf. Corollary IV.1), such as C does not affect XY given AB, etc. Thus, in the language of the present paper, the result of [30] can be stated as saying that given an embedding of the form $\mathscr{E}^{\text{jam}}$, and with A, B and C exogenous, the conditions **NS3′** are sufficient to prevent superluminal signalling by interventions on A, B and C.

A stronger claim is made by [31], that the conditions **NS3′** are necessary and sufficient for ensuring no superluminal signalling and no causal loops with such an embedding and they termed the correlations satisfying **NS3′** "relativistic causal correlations". Within the framework introduced in this paper, if interventions are also allowed on X, Y and Z, the sufficiency of **NS3′** for ruling out superluminal signalling in the space-time embedding $\mathscr{E}^{\text{jam}}$ does not hold. The reason is that there are causal models satisfying **NS3′** as well as having X affects Y. The latter involves intervention on a non-exogenous node X, and implies that Alice can signal to Bob. With an embedding of the form $\mathscr{E}^{\text{jam}}$, this is

[15] The conditions $\mathcal{A} \prec \mathcal{X}$, $\mathcal{B} \prec \mathcal{Y}$, $\mathcal{C} \prec \mathcal{Z}$ are more natural than the last three relations, but, as in [31], we allow for the possibility of instantaneous measurements here.

superluminal. This is also captured by our definition of compatibility (Definition V.7), according to which the relation X affects Y is not compatible with an embedding of the form $\mathscr{E}^{\text{jam}}$.

The claim of [31] is not just about the impossibility of superluminal signalling but also about ruling out causal loops. As we have seen, correlations satisfying (**NS3′**) that allow for jamming must be fine-tuned regardless of the causal structure. In fine-tuned causal models, several distinct classes of causal loops are possible, some of which are operationally undetectable (or hidden — cf. Definition VI.2) while others are operationally detectable but nevertheless do not lead to superluminal signalling as we show in the associated Letter [1] (see also Example VI.2). The former, by virtue of being operationally undetectable can never be ruled out only from the correlations (**NS3′** or otherwise) or affects relations. Operationally detectable causal loops require the consideration of non-trivial affects relations between sets of RVs, which are not detectable from the correlations alone. For instance **NS3′** cannot detect the existence of causal loops between outcome variables (e.g., X is a cause of Y and Y is a cause of X) since when a common cause is included, this common cause can explain the correlations. Therefore, the claim of [31] that **NS3′** rules out causal loops does not hold within our framework, even when restricting to the case where the settings are exogenous. More generally, in the absence of alternative frameworks for formalising these questions, it remains unclear how conditions on correlations such as **NS3′** could be necessary and sufficient for ruling out causal loops without further assumptions.

Furthermore, although we can rule out certain types of operationally detectable causal loops by demanding certain properties of the space-time embedding e.g., affects causal loops of Types 1–3 (cf. Lemmas VI.2 and VI.3), the absence of operationally detectable causal loops is not necessary to ensure no superluminal signalling (cf. Example VI.2). While **NS3′** is necessary to prevent superluminal signalling within the embedding $\mathscr{E}^{\text{jam}}$, it is not necessary to rule out affects causal loops in this embedding: it is possible to have an acyclic causal model over the settings A, B and C, and outcomes X, Y and Z that violates **NS3′** and leads to superluminal signalling in the embedding $\mathscr{E}^{\text{jam}}$. The implication that superluminal signalling implies causal loops holds within the theory of special relativity. Here we do not want to assume it, which allows us to consider more general relations between these principles and our framework can hence be used also in theories with a preferred frame for instance. A more detailed analysis of previous works such as [31] and [30] and the possibilities of superluminal signalling/causal loops in the jamming scenario are carried out in upcoming work [38].

VIII. SUMMARY AND CONCLUSIONS

We have developed a general mathematical framework for studying causality by clearly separating between operational and space-time related notions and characterising their compatibility. This has foundational relevance for understanding causality in quantum and more general theories, as well as practical applications for cryptography, information processing tasks in space-time and causal discovery. We have mainly focused on two notions of causality: the operational notion of causality defined through an extension of the causal modelling approach [2, 17] and relativistic causality which is associated with a space-time structure.

We formulated the operational notion of causality under minimal assumptions while allowing for causal influences to be fine-tuned, cyclic and mediated by latent non-classical systems. On the other hand, relativistic causality can be understood as the condition that "causal influences can propagate only from past to future in the space-time", where it has several implications such as "it is impossible to signal outside the future", "it is possible to signal everywhere in the future and nowhere else", "in Minkowski space-time it is impossible to have causal loops", and "it is impossible to broadcast classical information outside the future". Often one or more of these implications are taken in isolation to represent the condition of relativistic causality. Within the theory of special relativity these are related (e.g., the possibility of superluminal signalling leads to causal loops), but without assuming relativity they may not be and hence need to be independently formalised.

Within our framework we have formalised several of the above concepts and shown that these are distinct conditions in general. Our compatibility condition (Definition V.7) ensures that a causal model does not lead to signalling outside the future when embedded in a space-time structure. An alternative compatibility condition discussed in Section V D captures the idea of broadcasting classical variables within the space-time future. Cyclic causal models involve causal loops and when embedded in space-time, as described in Section V B, these allow for causal influences going in both directions between two distinct space-time points. Thus, the embedded cyclic causal structure can be understood as a closed time-like curve (CTC). Applying this framework, we have shown in the associated Letter [1] that it is mathematically possible to have such CTCs in Minkowski space-time, and that their existence can be operationally detected without leading to superluminal signalling. This establishes that no superluminal signalling and no causal loops/closed timelike curves are in general different conditions. In the present paper, we have gone beyond this particular example and identified several different classes of operationally detectable causal loops (or affects causal loops) in our framework and characterised properties of their space-time embeddings. Should one such

operational detection be made (which we do not expect) it would certify the physical existence of retro-causation. Such constructions are possible because our framework does not require all causal influences to respect the partial order of the space-time but only that signalling possibilities are constrained by the space-time.

In particular, this work also serves as the first causal modelling framework for a class of post-quantum theories ("jamming theories") previously proposed in the literature [30, 31] which are known to be more general than standard GPTs. Previously, these theories have been analysed focusing on the correlations they generate, but a causal modelling framework enables us to systematically study the effect of active interventions on arbitrary physical systems in such theories which provides more information about the underlying causal structure than correlations alone. Using this, we analysed previous claims regarding the compatibility of such theories with principles such as no superluminal signalling and no causal loops, which suggests that these claims cannot hold without further assumptions. In future work [38] we apply the framework developed here to such post-quantum jamming scenarios to characterise the signalling possibilities and new properties of theories admitting such correlations.

To allow us to deal with fine-tuned causal structures, we introduced higher-order affects relations. Our results show these to be a useful tool for inferring causation in the presence of fine-tuning that also has operational meaning in terms of signalling through joint interventions on multiple systems. When a particular phenomenon has two possible causal explanations, one of which is fine-tuned, the fine-tuned explanation is often considered undesirable because it usually more complicated and involves features that cannot be operationally verified. Fine-tuning complicates causal reasoning and the majority of the literature on causal models typically assumes the absence of fine-tuning. Explanations of quantum correlations in terms of classical causal models are typically rejected as such explanations involve fine-tuning [11], and instead faithful explanations in terms of quantum causal models are often preferred. On the other hand, fine-tuning occurs in many cryptographic scenarios, as well as jamming correlations (cf. Proposition III.1, Example IV.6). Using a causal modelling approach allows for a clear distinction to be made between undesirable and potentially useful forms of fine-tuning. The former correspond to causal influences that can never be operationally detected while the latter can be operationally detected by considering more general, joint interventions on sets of random variables. In this work, we have shown that several causal modelling concepts that are equivalent in the absence of fine-tuning, become distinct concepts in the presence of fine-tuning. We presented several technical results relating these various concepts in general fine-tuned and cyclic causal models with latent non-classical causes, which can have useful applications for the causal discovery problem in the presence of fine-tuning.

In our framework, we have modelled space-time as a discrete partially ordered set on which we embedded a separate operational causal model, and considered compatibility between the two notions. There are two different ways in which this space-time can be interpreted. One is to regard it as a fundamental background on which physics given by the causal model is embedded, such that every observed node in the model can be associated with a location in the space-time. Then our results tell us that the absence of signalling outside the future when the causal model is embedded in the space-time, does not allow us to identify the space-time order relation with the causal order, and that these are distinct concepts. A second interpretation is to understand the space-time order as an emergent property of the physics given by the causal model. For instance, our compatibility condition could be interpreted as a way to infer which space-time orders could occur alongside the operational predictions of the model, if we consider the direction of signalling in the model to constrain the order relations of the space-time. Our works (the present paper and [1]) show that even in cyclic causal models, it can be possible to single out a preferred direction (namely the direction of signalling) from the operational predictions of the model, while at the same time certifying that the underlying causal model is cyclic.

The present work focuses on the signalling possibilities allowed by the causal model, rather than the strength of signalling or correlations, even though the framework developed here can in principle model both. In [52], the strength of correlations was considered as a way to capture properties such as distance that are associated with an underlying space-time, with the hope that space-time can be seen as emergent. In approaches to quantum gravity, such as causal set theory [53–55], an active line of research is to derive geometric properties of a continuum space-time from order-theoretic properties of discrete graphs that capture the causal relations of the space-time. The present work, along with a related follow-up work [29] suggest a possible direction of inquiry for connecting the research on non-classical causal models with such approaches to quantum gravity, and we leave these interesting directions for future work.

To summarise, our results highlight the importance of separating a) operational and space-time related notions of causality b) correlation, causation and signalling (by considering interventions) and c) distinct notions of causality within the operational/space-time categories mentioned in a).

IX. OPEN QUESTIONS

The work presented here provides a platform for analysing a number of problems in quantum foundations and causality in a new light. We discussed specific interesting and open questions related to the characterisation of causal

loops within our framework in Remark VI.2. Here we place our work within a broader context and discuss the associated open questions.

a. Other notions of causality: While this work elucidates the relationships between a number of different notions of causality, there are many more that may be considered. For instance, another operational notion of causality is that of process terminality [56] which says that discarding all the outputs of a causal process is equivalent to discarding the process. Further, approaches such as the process matrix framework [32] aim to formulate causality more generally in the absence of a fixed background space-time (which we have assumed here). Other setup assumptions in these approaches mean that, for instance, post-quantum jamming scenarios cannot be modelled.[16] Here several conditions such as causal orderedness, causal separability, satisfaction of causal inequalities have been proposed, which serve as causality criteria under different assumptions. The precise relationships between all these notions of causality, their operational meaning and implications for the physics of information processing remains open. In a related work involving one of the authors [29], the present approach of disentangling operational and space-time notions of causation and characterising their compatibility, is applied to operational scenarios described by the process matrix framework [32]. There, further connections between indefinite and cyclic causation are established in quantum scenarios and a more general class of space-time embeddings is considered that allows for spacetime embeddings of quantum systems where the systems are not nonlocalised to a single space-time location but may possibly be delocalised over a space-time region. These results (along with previous works such as [28]) relating indefinite causation to definite cyclic causation indicate that cyclic and non-classical causal models can have applications also to scenarios where a background space-time structure is not assumed. Although we do not consider it here, our framework can also be used to analyse frame-dependent notions of causality associated with Minkowski space-time (e.g., whether compatibility and other properties of an embedding can depend on the choice of classical reference frame).[17] Another intriguing prospect for future research would be to consider compatibility between operational causal models and space-time related information in more exotic regimes where a global space-time structure may not exist but agents infer space-time information using their local (possibly quantum) reference frames [33, 57, 58].

b. Affects relations and d-separation: We use affects relations (Definition IV.3), based on the notion of interventions, to distinguish between correlation and causation. In acyclic causal structures [2, 17] and in classical cyclic causal structures [25], existing frameworks prescribe how the post-intervention distribution can be calculated from the observed distribution and/or the underlying causal mechanisms. In non-classical cyclic causal structures, such a characterisation is not available. In Section IV, we used the d-separation condition (Definition IV.1) on the observed distribution to obtain a partial characterisation which suffices for the current purpose, but this does not fully specify the post-intervention distribution. In Appendix C, we outline a possible method for obtaining the post-intervention distribution given the underlying causal mechanisms. [Although this method may not always recover the d-separation condition IV.1, this does not impact the results of this paper.] Generalising our framework to also include non-classical cyclic causal models that do not obey the d-separation condition, by using the causal mechanisms as primitives would allow our results regarding space-time compatibility and affects loops to be applied to this larger class of models. This would provide a general framework for causally modelling fine-tuned and cyclic non-classical causal models such that any post-intervention scenario can be completely specified by the original causal model.[18] Another observation made in Appendix C is that the presence of causal loops could allow us to distinguish between faithful, non-classical explanations and unfaithful classical explanations (e.g., using non-local hidden variables) of quantum correlations, which cannot be operationally distinguished otherwise. This suggests that it might be possible to operationally distinguish hidden variable interpretations of quantum theory such as Bohmian mechanics from inherently "quantum" interpretations, in the presence of causal loops. Formalising this observation would be another interesting line of investigation.

---

[16] The process framework assumes a tensor product structure between the local operations of various parties, and once communication between parties is forbidden, the framework can only produce correlations compatible with standard no-signalling theories and not the relaxed no-signalling conditions of [31] that permit jamming.

[17] For example, one can consider a different partially ordered set to represent space-time structure from the perspective of different frames, such that classical frame transformations such as Lorentz transformations could be viewed as invertible maps between these partially ordered sets.

[18] In the current framework, the causal model is defined in terms of the observed distribution and therefore not all affects relations can be deduced from the model's specification. When characterised instead in terms of the causal mechanisms, the affects relations should become deducible.

c. Causal loops and paradoxes: In this work we have considered space-time to be modelled by a partial order. The theory of general relativity allows for the possibility of more exotic space-time structures. These possibilities led to investigations of closed time-like curves and there are mathematical models of CTCs that are logically consistent and do not lead to time travel paradoxes [59–63]. Two inequivalent models have been developed to make sense of information flow in the presence of CTCs, Deutsch's CTCs (DCTCs) [59] and post-selected CTCs (PCTCs) [60–63]. DCTCs and PCTCs are known to have different amounts of computational power [63, 64] and to provide different resolutions to the grandfather and unsolved theorem paradoxes [63]. In our framework, grandfather-type paradoxes are forbidden by the assumption that a valid joint probability distribution observed variables exists, which implies that the underlying causal mechanisms (e.g., functional equations in the classical case) must be mutually consistent.[19]

The unproved theorem paradox on the other hand is not ruled out in the current framework, and can depend on how the framework is further instantiated with causal mechanisms. For example, in classical cyclic causal models, an assumption regarding the unique solvability of the underlying functional dependences is often considered. In particular, this could be seen as the requirement that any information involved in a loop (such as the unproved theorem) must be fully and uniquely determined by the mechanisms of the causal model thereby eliminating the paradox of a proof that "came from nowhere". An analogous condition on the causal model for ruling out the unproved theorem paradox in the quantum case is far from clear, since the causal mechanisms in this case are not deterministic functional equations. These questions can be explored within a full formalisation of our framework in terms of causal mechanisms, along the lines discussed in the previous paragraph (and Appendix C). It is interesting to consider whether there are connections between the CTCs that can be embedded in Minkowski space-time without superluminal signalling (such as Example VI.2) and DCTCs or PCTCs, or which physical principles rule out such CTCs.

d. Causality in time-symmetric formulations of quantum theory: Unitary quantum mechanics is time symmetric while operational quantum theory has the possibility of irreversible measurements. There are several proposals for modelling quantum and more general experiments in a time symmetric manner while still making operational predictions and retrodictions about measurements [65–69]. Predictions and retrodictions indicate the direction of inference, not necessarily of causation and the role of causality (in terms of a causal modelling paradigm) is not fully understood in these frameworks. A notable approach for making operational statements in a time-symmetric setting is the two-time state formalism [65, 70, 71], which describes measurements on pre- and post-selected quantum states, where the former can be thought of as evolving forward-in-time and the latter, backward-in-time. It is interesting to consider how this time-symmetric approach can be modelled in a causal framework.

e. Causal discovery in the presence of fine-tuning: Causal discovery is the problem of inferring a fully or partially unknown causal structure from observed correlations, possibly combined with additional information about interventions. Fine-tuning makes this task more challenging and causal discovery algorithms typically assume that the underlying causal model is not fine-tuned [2, 3], even in cases where the underlying model is classical and has no unobserved nodes. Relaxations of this assumption have been considered where certain forms of fine-tuning (but not all) have been allowed [72]. Intuitively, use of higher-order affects relations appears useful for causal discovery in the presence of fine-tuning, and the examples of Section IV C show the usefulness of HO affects to distinguish between causal structures with the same correlations. We believe this deserves future exploration.

f. Indefinite space-time locations: In the present work, we have embedded causal models in a space-time structure by assigning a single space-time location to each observed system. More generally, we can have, both in theory and practice, systems whose space-time locations are associated with some classical or quantum uncertainty or protocols involving quantum systems that are delocalised over space and in time [29, 73–76]. It would therefore be of interest to generalise our methods to allow for such superpositions. In a related work [29], a method to do this for finite dimensional quantum systems in a discrete space-time (i.e., a partially ordered set as considered here) is proposed, which has applications for physically characterising so-called indefinite causal order processes [32].

---

[19] An example of a paradoxical scenario is a 2-cycle between binary variables X and Y where the influence $X \longrightarrow Y$ defines the functional dependence $Y = X$ and $Y \longrightarrow X$ gives the dependence $X = Y \oplus 1$. These equations are mutually inconsistent and there is no joint distribution P_{XY} compatible with these dependences.

Acknowledgments

A preliminary version of this work and upcoming work [38] is available in VV's PhD thesis [77] (Chapter 6). VV thanks Maarten Grothus for useful feedback on the framework. We also thank Mirjam Weilenmann and Lorenzo Maccone for insightful discussions. VV acknowledges support from the PhD scholarship of the Department of Mathematics, University of York and the ETH Postdoctoral Fellowship from ETH Zürich.

Appendix A: Identifying conditional independences and affects relations: Examples

Here we provide examples that better illustrate some of the definitions and rules of the framework. In particular, how one can deduce the conditional independences and affects relations in a given causal model. For this the following lemmas will be useful, these can be regarded as generalisations of Corollary IV.2 and Lemma IV.2 from the unconditional zeroth-order to the case of general conditional higher-order affects relations, and are proven in Appendix D 4.

Lemma A.1. *For any pairwise disjoint subsets X, Y, Z and W of the observed nodes S of a causal model, we have*

1. $(XZW \perp^d Y)_{\mathcal{G}_{\mathrm{do}(XZ)}} \Rightarrow XZ$ *does not affect* Y *given* W.
2. $(XZW \perp^d Y)_{\mathcal{G}_{\mathrm{do}(XZ)}} \Rightarrow X$ *does not affect* Y *given* $\{\mathrm{do}(Z), W\}$.

Lemma A.2. *For any pairwise disjoint subsets X, Y, Z and W of the observed nodes S of a causal model, we have*

1. $(XZW \not\perp\!\!\!\perp Y)_{\mathcal{G}_{\mathrm{do}(XZ)}} \Rightarrow XZ$ *affects* Y *given* W.
2. $(XZW \not\perp\!\!\!\perp Y)_{\mathcal{G}_{\mathrm{do}(XZ)}}$ and $(ZW \perp^d Y)_{\mathcal{G}_{\mathrm{do}(Z)}} \Rightarrow X$ affects Y *given* $\{\mathrm{do}(Z), W\}$.

We now summarise how one may use these results to deduce some of the conditional independences and affects relations from a given causal model.

- *Conditional independences:* Given a causal graph $\mathcal{G}$ with the set of observed nodes S, some of the conditional independences satisfied by the joint distribution P_S can be identified using Definition IV.1 i.e., by listing all the conditional independences implied by d-separation relations in $\mathcal{G}$. Further independences may be found if there are dashed arrows emanating from exogenous nodes, since $X \dashrightarrow Y$ implies X does not affect Y (by Definition IV.4) which implies $X \perp\!\!\!\perp Y$ if X is exogenous (cf. Corollary IV.1). Lemma IV.1 can also be used to list further independences not directly implied by d-separation in $\mathcal{G}$. There may still be more conditional independences in P_S that cannot be listed using the methods mentioned above. Since we allow for fine-tuning and latent systems, there could in principle be arbitrarily many independences in P, but those mentioned above are sufficient for compatibility with the causal model.
- *Affects relations:* The existence of an affects relation X affects Y given $\{\mathrm{do}(Z), W\}$ can be deduced by applying Lemma A.2 (for the zeroth-order case, X affects Y is deduced by applying Lemma IV.2). The non-affects relation X does not affect Y given $\{\mathrm{do}(Z), W\}$ can be deduced by applying Lemma A.1 and also the contrapositives of Lemmas IV.3, IV.4 and IV.5. For example, $(X \perp^d YW)_{\mathcal{G}_{\mathrm{do}(X)}}$ implies that X does not affect Y given $\{\mathrm{do}(Z), W\}$ by the second statement of Lemma IV.3. In the zeroth-order case, the non-affects relation X does not affect Y is deduced by applying Corollary IV.2. The direction of the lemmas is important to note here, for instance the converse of Lemma IV.2 cannot be used to deduce that X does not affect Y since this implication does not hold, unless X is exogenous (cf. non-implication 2 of Figure 7 and Corollary IV.1). Therefore one can check for non-independences and d-separations in the post-intervention causal model to identify affects and non-affects relations respectively. One may be able derive further results of this sort or exploit structural aspects of particular causal models to derive additional independences and affects relations. Again, due to fine-tuning and latent systems, in general, this identification may not be exhaustive even after this is done. However, in the case of causal models with no latent nodes (which are by definition, classical), it would indeed be exhaustive, as explained in Remark IV.3.

In case some or all of the causal mechanisms are also given in addition to the observed distributions, it may be possible to identify further independences and affects relations in the model. We now apply these rules to specific examples where it is possible to deduce all the conditional independences and affects relations involved, these are tabulated in Figure 13 for 3 out of 4 of the examples considered here, all of which correspond to fine-tuned causal models. The fourth example corresponds to a faithful but cyclic causal model, and therefore the d-separation condition IV.1 and zeroth-order affects relations (given by the causal arrows themselves) completely characterise the scenario.

(S_1,S_2,S_3)	$(S_1 \perp^d S_2 \mid S_3)$?			$(S_1 \perp\!\!\!\perp S_2 \mid S_3)$?			S_1 affects S_2 given do(S_3)?			S_1 affects S_2 given S_3?		
	jamming	collider	ACL4	jamming	collider	ACL4	jamming	collider	ACL4	jamming	collider	ACL4
$(A, B, \varnothing)$	×	×	×	✓	✓	✓	×	×	×	×	×	×
$(B, A, \varnothing)$	×	×	×	✓	✓	✓	×	×	×	×	×	×
$(B, C, \varnothing)$	×	×	×	✓	✓	✓	×	×	×	×	×	×
$(C, B, \varnothing)$	×	×	×	✓	✓	✓	×	×	×	×	×	×
$(C, A, \varnothing)$	×	✓	×	✓	✓	✓	×	×	×	×	×	×
$(A, C, \varnothing)$	×	✓	×	✓	✓	✓	×	×	×	×	×	×
$(A, BC, \varnothing)$	×	×	×	×	×	×	×	✓	✓	×	✓	✓
$(BC, A, \varnothing)$	×	×	×	×	×	×	×	×	×	×	×	×
$(B,AC, \varnothing)$	×	×	×	×	×	×	✓	×	✓	✓	×	✓
$(AC, B, \varnothing)$	×	×	×	×	×	×	×	✓	✓	×	✓	✓
$(C,AB, \varnothing)$	×	×	×	×	×	×	×	✓	✓	×	✓	✓
$(AB, C, \varnothing)$	×	×	×	×	×	×	×	×	×	×	×	×
(A, B, C)	×	×	×	×	×	×	×	✓	✓	×	✓	✓
(B, A, C)	×	×	×	×	×	×	×	×	×	✓	×	✓
(A, C, B)	×	×	×	×	×	×	×	×	×	×	✓	✓
(C, A, B)	×	×	×	×	×	×	×	×	×	×	✓	✓
(B, C, A)	×	×	×	×	×	×	×	×	×	✓	×	✓
(C, B, A)	×	×	×	×	×	×	×	✓	✓	×	✓	✓

FIG. 13: Table of all possible d-separations, conditional independences and affects relations for jamming, fine-tuned collider and Type 4 affects causal loop examples (Sections A 1, A 2, A 3), all of which involve the three observed RVs A, B and C. All affects relations, when they do exist, are irreducible.

1. Jamming (Figure 9a)

In the jamming causal structure $\mathcal{G}^{\text{jam}}$ of Figure 9a and Example IV.6, Definition IV.1 does not impose any conditional independences on the observed distribution P_{ABC} since Λ is unobserved.[20] However, from Definition IV.4 of dashed arrows we know that B affects neither A nor C individually and we are given that B affects AC. Using the exogeneity of B (cf. Corollary IV.1), this implies the independences $A \perp\!\!\!\perp B$ and $C \perp\!\!\!\perp B$ and the non-independence $B \not\perp\!\!\!\perp AC$ in $\mathcal{G}^{\text{jam}}$. Now, consider an intervention on A. The post-intervention causal structure $\mathcal{G}^{\text{jam}}_{\text{do}(A)}$ only has the edges $B \dashrightarrow C$ and $\Lambda \rightsquigarrow C$ (along with $I_A \longrightarrow A$ of course). The d-separation $(A \perp^d C)_{\mathcal{G}^{\text{jam}}_{\text{do}(A)}}$ implies the independence $(A \perp\!\!\!\perp C)_{\mathcal{G}^{\text{jam}}_{\text{do}(A)}}$ (Definition IV.1) and also that A does not affect C (Corollary IV.2). Similarly, we can derive the relations C does not affect A, A does not affect BC, C does not affect AB and AC does not affect B. Combined with the lack of any zeroth-order affects relations between any two of the RVs, this implies that there are no affects relations of order 1 or higher (by the contrapositive of Lemma IV.4). Further, using Lemma IV.1 and the exogeneity of B, we can derive AB does not affect C as follows. In the causal structure $\mathcal{G}^{\text{jam}}_{\text{do}(AB)}$, A is d-separated from B and C, while B and C are independent of each other due to the exogeneity of B and the dashed arrow connecting them. Using the lemma, this gives $(AB \perp\!\!\!\perp C)_{\mathcal{G}^{\text{jam}}_{\text{do}(AB)}}$ which can be explicitly written as $P_{\mathcal{G}^{\text{jam}}_{\text{do}(AB)}}(C|A,B) = P_{\mathcal{G}^{\text{jam}}_{\text{do}(AB)}}(C)$. The right hand side can be simplified in the following two steps. Firstly as $P_{\mathcal{G}^{\text{jam}}_{\text{do}(AB)}}(C) = P_{\mathcal{G}^{\text{jam}}_{\text{do}(A)}}(C)$ noting that $\mathcal{G}^{\text{jam}}_{\text{do}(AB)}$ and $\mathcal{G}^{\text{jam}}_{\text{do}(A)}$ are effectively the same graph due to the exogeneity of B. Then the d-separation $(A \perp^d C)_{\mathcal{G}^{\text{jam}}_{\text{do}(A)}}$ implies the independence $P_{\mathcal{G}^{\text{jam}}_{\text{do}(A)}}(C|A) = P_{\mathcal{G}^{\text{jam}}_{\text{do}(A)}}(C)$, which along with A does not affect C (as noted earlier) gives $P_{\mathcal{G}^{\text{jam}}_{\text{do}(A)}}(C) = P^{\text{jam}}_{\mathcal{G}}(C)$. Putting this together, we get $P_{\mathcal{G}^{\text{jam}}_{\text{do}(AB)}}(C|A,B) = P^{\text{jam}}_{\mathcal{G}}(C)$ i.e., $\{A.B\}$ does not affect C. Similarly, one can obtain BC does not affect A. All these are summarised in Figure 13.

[20] If Λ in Figure 9a were observed, A and C would be d-separated given $\{B,\Lambda\}$ and we would have the conditional independence $P_{AC|B\Lambda} = P_{A|B\Lambda}P_{C|B\Lambda}$.

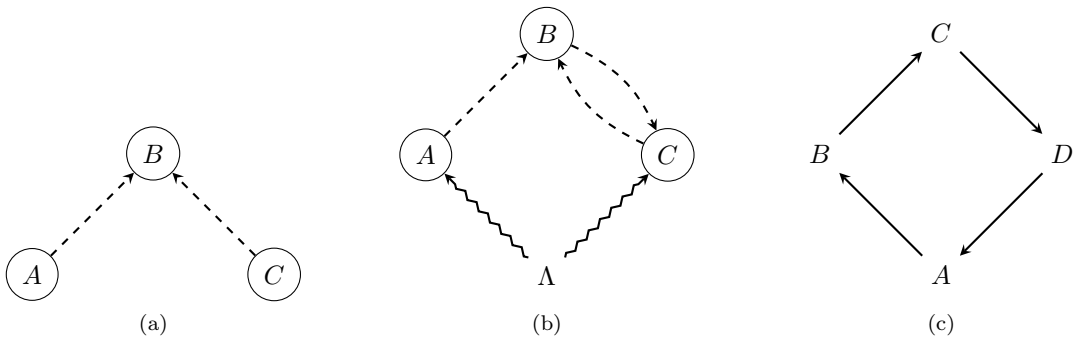


FIG. 14: **Some fine-tuned and/or cyclic causal structures:** **(a)** A fine-tuned collider **(b)** A Type 4 affects causal loop **(c)** A Type 1 affects causal loop

2. Fine-tuned collider (Figure 14a)

In the causal structure of Figure 14a, the independence $A \perp\!\!\!\perp C$ follows from the d-separation condition (Definition IV.1), while $A \perp\!\!\!\perp B$ and $C \perp\!\!\!\perp B$ follow from the dashed arrow structure. These are the same independences as the case in the previous example of jamming with unobserved Λ (where $A \perp\!\!\!\perp C$ was an additional independence in the jamming example but follows from d-separation in this case). Thus the distribution P_{ABC} from Example IV.6 is compatible with both the jamming (Figure 9a) as well the fine-tuned collider (Figure 14a) causal structures.[21] However interventions on the two causal structures yield different results. We have AC affects B for the fine-tuned collider (since AC consists of exogenous nodes and is correlated with B) but not for the jamming case. We also have A affects BC and C affects AB for the fine-tuned collider even though A and C do not individually affect B due to the dashed arrow structure. This follows from the exogeneity of A and C and the joint correlations $A = B \oplus C$. Further, AB does not affect C since these sets become d-separated upon intervention on AB and by a similar reasoning, BC does not affect A and B does not affect AC (in contrast with the jamming case where B affects AC). As for higher-order affects relations, A affects B given $\mathrm{do}(C)$ and C affects B given $\mathrm{do}(A)$ are the only ones, and these follow by applying Lemma A.2 to this example. Again, these conclusions are summarised in Figure 13.

3. A Type 4 affects causal loop ([1], Figure 14b)

Example VI.2 outlined a cyclic causal model proposed in [1] for demonstrating the mathematical possibility of compatibly embedding affects causal loops in Minkowski space-time. Here, we reproduce further details of this model and discuss its properties it in the context of the more general framework developed in the present paper. Consider the cyclic causal structure $\mathcal{G}^{\mathrm{ACL4}}$ of Figure 14b along with the following classical causal mechanisms where all 4 variables are taken to be binary: $A = \Lambda$, $B = A \oplus C$, $C = B \oplus \Lambda$, where the exogenous variable Λ is uniformly distributed. One can check that the distribution P_{ABC} obtained through these mechanisms would be the same as that of the jamming as well as the fine-tuned collider examples above, but the affects relations differ from those of these examples and instead correspond to those of Example VI.2 which is an affects causal loop of Type 4. To obtain these affects relations, first note that in the causal model of $\mathcal{G}^{\mathrm{ACL4}}_{\mathrm{do}(A)}$, Λ is no longer a parent of A, but using the remaining causal mechanisms $B = A \oplus C$ and $C = B \oplus \Lambda$ (which remain the same), we can still obtain $A = \Lambda$. Therefore the intervention on A does not change the observed distribution and A and B continue to be independent in $\mathcal{G}^{\mathrm{ACL4}}$ as well as $\mathcal{G}^{\mathrm{ACL4}}_{\mathrm{do}(A)}$, and in both graphs the marginal distributions over A, B and C are uniform, which gives A does not affect B. On the other hand, B does not affect A can be established simply from the d-separation $(B \perp^d A)_{\mathcal{G}^{\mathrm{ACL4}}_{\mathrm{do}(B)}}$. In the causal model of $\mathcal{G}^{\mathrm{ACL4}}_{\mathrm{do}(C)}$, neither B nor Λ are parents of C but the remaining mechanisms $A = \Lambda$ and $B = A \oplus C$ give $C = B \oplus \Lambda$. Again, the observed distribution here is the same as the pre-intervention distribution, which gives C does not affect B. By a similar argument, B does not affect C can also be established. Further, we have both B affects AC (as in the jamming case) *and* AC affects B (as in the fine-tuned collider) since $P_{\mathcal{G}^{\mathrm{ACL4}}_{\mathrm{do}}(AC)}(B|A,C)$ and $P_{\mathcal{G}^{\mathrm{ACL4}}_{\mathrm{do}}(B)}(A,C|B)$

[21] Note that this is essentially the one-time pad example from earlier.

are deterministic while $P_{\mathcal{G}^{\text{ACL4}}}(B)$ and $P_{\mathcal{G}^{\text{ACL4}}}(A, C)$ are uniform. We also have A affects BC and C affects AB as in the fine-tuned collider, which can be verified using the causal mechanisms given.[22] As in the jamming case, we also get AB does not affect C and BC does not affect A. The higher-order affects relations here are identical to the previous example, and obtained in a similar manner and these are given in Figure 13. Furthermore, even though this corresponds to an affects causal loop the existence of which can be operationally certified and this causal model admits a non-trivial and compatible embedding in Minkowski space-time as explained i.e., it does not lead to superluminal signalling [1] (see also Example VI.2).

4. A Type 1 affects causal loop (Figure 14c)

Consider a faithful causal model associated with the cyclic causal structure of Figure 14c. Here all 4 nodes are observed and hence classical. The faithfulness implies that the only conditional independences are those implied by d-separation. Since $B \perp^d D|AC$ and $A \perp^d C|BD$ are the only d-separations, $B \perp\!\!\!\perp D|AC$ and $A \perp\!\!\!\perp C|BD$ are the only conditional independences. Further, the faithfulness also implies that every causal arrow is associated with an affects relation, which is also reflected in the fact that all arrows are solid arrows, and we have A affects B, B affects C, C affects D and D affects A, which forms an affects causal loop of Type 1 (there is an affects relation is both direction between every pair of RVs). One can easily check that the only irreducible affects relations are the zeroth-order affects relations between single RVs, as one would expect for faithful causal models. To further illustrate the kind of causal loops allowed in this framework, consider the pairwise correlations $A = B$, $B = C$, $C = D$ and $D \neq A$. Since this system of equations has no solutions, there exists no joint distribution P_{ABCD} from which the pairwise marginals producing these correlations can be obtained. Such examples correspond to grandfather type paradoxes and cannot be modelled in frameworks that demand the existence of a valid joint probability distribution over all variables involved in a causal loop. On the other hand, examples of solid arrow directed cycles where the functional dependences of the loop variables admit solutions such as $A = B = C = D$ (with any probability) or the examples considered in [43] for other cyclic causal structures can be modelled in our framework. Additionally, there can also be Type 1 and Type 2 affects causal loops that do not involve any solid arrows, for example through a concatenation of structures such as that of Figure 9b. We discuss causal loops in more detail in Appendix C, also in the case of quantum causal structures.

Appendix B: Further classes of affects causal loops and their space-time embeddings

As motivated in the main text (see the paragraph after Definition VI.8), we can consider further classes of affects causal loops that are distinct from ACL1, ..., ACL6. The intuition is that the chain of irreducible affects relations involved in these previous definitions is such that for any two adjacent affects relations in the chain the second set of the first is contained in the first set of the second. Relaxing this containment condition can lead to a violation of Theorem VI.1, as also explained in the main text. So we can consider relaxing this condition and only requiring a non-trivial intersection between the sets (which would make the chain "incomplete"), as long as we include additional conditions on the affects relations that will again guarantee cyclicity of the causal structure. Here we propose four more classes of affects causal loops ACL7, ACL8, ACL9 and ACL10 based on this idea, illustrate them with examples which also show that there can in general be more classes of affects causal loops even beyond these.

Definition B.1 (Affects causal loops, Type 7 (ACL7))**.** A set of affects relations $\mathscr{A}$ is said to contain a Type 7 affects causal loop if the following conditions are satisfied

1. There exist disjoint sets of RVs S_1 and S_2 such that S_1 affects S_2 belongs to $\mathscr{A}$ and is irreducible.
2. There exists a chain of irreducible affects relations (possibly incomplete) $\mathscr{C}_{s_2}$ from some subset $s_2 \subseteq S_2$ to S_1 i.e., there exists sets of RVs $s_2 \subseteq S_2'$, S_3, S_3', ... S_n, S_n', $s_1 \subseteq S_1$ such that $\{S_2'$ affects S_3, S_3' affects S_4, ... ,S_{n-1}' affects S_n, S_n' affects $s_1\} \subseteq \mathscr{A}$, where all the affects relations are irreducible, every pair of sets connected by an affects relation is disjoint, $S_i \bigcap S_i' \neq \varnothing$ for all $i \in \{3, \ldots, n\}$ and $S_2 \bigcap S_2' = s_2$. Each pair (S_i, S_i') such that $S_i \notin S_i'$ for $i \in \{2, \ldots, n\}$ is called an incomplete node of the affects chain $\mathscr{C}_{s_2}$, a complete affects chain has no incomplete nodes.

[22] Note that in the absence of the causal mechanisms, many of the affects/non-affects relations may not be identifiable. For example, to deduce that AB does not affect C in the jamming case, we used Lemma IV.1 along with the fact that B was exogenous in $\mathcal{G}^{\text{jam}}$. However, the same argument cannot be applied here since B is not exogenous in $\mathcal{G}^{\text{ACL4}}$.

3. For each affects chain $\mathscr{C}_{s_2}$ that connects the subset s_2 of S_2 back to S_1 as above, and each incomplete node (S_i, S_i') in $\mathscr{C}_{s_2}$ (for $i \in \{2, \ldots, n\}$), there exists a complete affects chain $\mathscr{D}^{\mathscr{C}}_{s_2}$ in $\mathscr{A}$ from $S_i \backslash (S_i \cap S_i')$ to S_i.

Definition B.2 (Affects causal loops, Type 8 (ACL8))**.** A set of affects relations $\mathscr{A}$ is said to contain a Type 8 affects causal loop if the following conditions are satisfied

1. There exist disjoint sets of RVs S_1 and S_2 such that S_1 affects S_2 belongs to $\mathscr{A}$ and is irreducible.

2. For each element $e_2 \in S_2$, there exists a chain of irreducible affects relations (possibly incomplete) $\mathscr{C}_{e_2}$ that connects it back to S_1 i.e., for each e_2, there exists sets of RVs $e_2 \in S_2'$, S_3, S_3', … S_n S_n', $s_1 \subseteq S_1$ such that $\{S_2'$ affects S_3, S_3' affects S_4, … ,S_{n-1}' affects S_n, S_n' affects$s_1\} \subseteq \mathscr{A}$, where all the affects relations are irreducible, every pair of sets connected by an affects relation is disjoint, $S_i \cap S_i' \neq \varnothing$ for all $i \in \{3, \ldots, n\}$ and $S_2 \cap S_2' = e_2$.

3. For each element $e_2 \in S_2$, each affects chain $\mathscr{C}_{e_2}$ that connects it back to S_1 as above, and each incomplete node (S_i, S_i') in $\mathscr{C}_{e_2}$ (for $i \in \{2, \ldots, n\}$), there exists a complete affects chain $\mathscr{D}^{\mathscr{C}}_{e_2}$ in $\mathscr{A}$ from $S_i \backslash (S_i \cap S_i')$ to S_i.

The following theorem (proven in Appendix D 5) generalises Theorem VI.1 to ACL7 and ACL8, and justifies categorising them as affects causal loops.

Theorem B.1. *Any set of affects relations $\mathscr{A}$ containing an affects causal loops of Type 7 or Type 8 can only arise from a causal model over a cyclic causal structure.*

More generally, for a given affects chain $\mathscr{C}_{s_2}$ in Definition B.1, and an incomplete node (S_i, S_i') in $\mathscr{C}_{s_2}$, instead of a single complete affects chain $\mathscr{D}^{\mathscr{C}}_{s_2}$ we could consider a set of incomplete affects chains that serve the same purpose and for which an analogous theorem holds. For example, for each incomplete node (S_i, S_i') of $\mathscr{C}_{s_2}$, there can exist an incomplete affects chain $\mathscr{D}^{\mathscr{C}}_{s_2}$ in $\mathscr{A}$ from $S_i \backslash (S_i \cap S_i')$ to S_i, such that for each incomplete node (R_j, R_j') of $\mathscr{D}^{\mathscr{C}}_{s_2}$, there exists another complete affects chain in $\mathscr{A}$ from $R_j \backslash (R_j \cap R_j')$ to R_j. This could go on recursively for arbitrarily many chains depending on the number of RVs appearing in $\mathscr{A}$. This recursive definition defines yet another class ACL9, and an analogous recursive version of ACL8 would define another class ACL10. Theorem B.1 for ACL9 and ACL10 follows through similar arguments, so we note this point without proof. We illustrate these new classes with some examples, along with an example to show that these (ACL1-ACL10) do not cover all possible affects causal loops.

Example B.1 (A Type 7 affects causal loop)**.** Consider the set of irreducible affects relations $\mathscr{A} = \{X$ affects Y, Y affects AB, A affects X, C affects AB, B affects $C\}$. One can check that $\mathscr{A}$ does not contain affects causal loops of Types 1 to 6, since no affects relation in $\mathscr{A}$ is such that every element of the second set has a complete affects chain leading it back to the first set. It however contains at least one Type 7 affects loop. For the affects relation Y affects AB, we have the incomplete chain $\mathscr{C}_A = \{A$ affects X, X affects $Y\}$ that connects A to Y with the incomplete node $(S_2 = AB, S_2' = A)$. Then $S_2 \backslash (S_2 \cap S_2') = \{B\}$ and we have the complete affects chain $\mathscr{D}^{\mathscr{C}}_A = \{B$ affects C, C affects $AB\}$ that connects $S_2 \backslash (S_2 \cap S_2')$ to S_2 as required.

Example B.2 (A Type 9 affects causal loop)**.** Consider the set of irreducible affects relations $\mathscr{A} = \{X$ affects Y, Y affects AB, A affects X, C affects AB, B affects CD, D affects E, E affects $CD\}$. This set is similar to the previous example, but does not contain a Type 7 causal loop (or ACLs of any lower types). It does contain a Type 9. For the affects relation Y affects AB, there is an incomplete chain $\mathscr{C}_A = \{A$ affects X, X affects $Y\}$ that connects A to Y as before. However, we have no complete chains from $S_2 \backslash (S_2 \cap S_2') = \{B\}$ to $S_2 = AB$ as before, only the incomplete chain $\mathscr{D}^{\mathscr{C}}_A = \{B$ affects CD, C affects $AB\}$. The incomplete node (R_j, R_j') of $\mathscr{D}^{\mathscr{C}}_A$ is $(R_j = CD, R_j' = C)$ and we have a complete affects chain $\{D$ affects E, E affects $CD\}$ from $R_j \backslash (R_j \cap R_j') = \{D\}$ to R_j.

Example B.3 (An affects causal loop not covered by Types 1 to 10)**.** Consider the set of irreducible affects relations $\mathscr{A} = \{X$ affects$Y$, Y affects AB, A affects X, C affects AB, B affects CD, BD affects $AC\}$. With some effort, one can see that $\mathscr{A}$ does not contain affects causal loops of Types 1–10. It nevertheless implies cyclicity, as follows. Applying Corollary IV.3, we have that X is a cause of Y, Y is either a cause of A or of B and A is a cause of X. If Y is a cause of A, we already have a directed cycle, so consider the case where Y is a cause of B. Using the remaining affects relations, we have C is a cause of either A or B, B is a cause of either C or D. Irrespective of whether C is a cause of either A or B, if B is a cause of C, we would have a directed cycle, so we must take B to be a cause of D to avoid this. The last affects relation implies that B is either a cause of A or of C. Irrespective of the choice here and the choice of whether C is a cause of A or of B, we can verify that there will always be a directed cycle. Hence this set of affects relations is an affects causal loop that is not of a previously defined Type.

Consider now, the space-time embedding for the affects relations of Example B.1 in Minkowski space-time. Imposing **compat** (Definition V.7) on the affects relations $\mathscr{A} = \{X$ affects Y, Y affects AB, A affects X, C affects AB, B affects $C\}$ implies that $\mathcal{Y}$ must contain the joint inclusive future of $\mathcal{A}$ and $\mathcal{B}$ but $\mathcal{A}$ is in the past of $\mathcal{X}$ which is in the

past of $\mathcal{Y}$. The only way this can be satisfied is if $\mathcal{B}$ is in the future of $\mathcal{A}$ such that the joint inclusive future of $\mathcal{A}$ and $\mathcal{B}$ coincides with the inclusive future of $\mathcal{B}$. The last two affects relations then imply that $\mathcal{B}$ and $\mathcal{C}$ must be embedded at the same location and since we have B affects C, this embedding is trivial. This implies that there is no non-trivial and compatible embedding of these affects relations in Minkowski space-time. In other words, the absence of affects causal loops of Types 1-6 does not guarantee the existence of a non-trivial and compatible space-time embedding. The presence of Type 3 and above ACLs does not rule out the existence of such an embedding as we have seen in Example VI.2, in contrast to the case of Type 1 and 2 ACLs (Lemma VI.2). This suggests that for each individual Type i of affects causal loops other than Types 1 and 2, the existence of a non-trivial and compatible space-time embedding is neither necessary nor sufficient for there to be no affects causal loops of that Type. By Lemma VI.3, for Type 3, the existence of a non-degenerate and compatible space-time embedding is sufficient but not necessary to rule out ACL3.

Appendix C: Do-conditionals from causal mechanisms in quantum cyclic causal models

In Section IV we outlined how interventions and do-conditionals (i.e., the post intervention distribution) are defined in our framework, and Theorem IV.1 provides some conditions under which the post and pre intervention distributions can be related. Ideally though, one would expect that it should be possible to fully specify the post-intervention distribution if we are given all the underlying causal mechanisms involved in the causal structure. For example, in the classical case, the structural equations of the causal model [2] provide these causal mechanisms. Here for each node X in the causal structure, the dependence of X on its parents $\text{par}(X)$ corresponds to a stochastic map, which can be written in terms of a deterministic function $X = f_X(\text{par}(X), E_X)$ by including an additional exogenous random variable E_X for each node X. This is called a *structural equation*. If the structural equations for all the nodes and the distributions of the parentless nodes are known, then the complete post-intervention distribution can be calculated. This has been shown to be the case for classical cyclic causal models in [25]. An intervention $\text{do}(x)$ on X, corresponds to updating the structural equation for X to $X = x$ while keeping the remaining structural equations the same. Another important result for the classical case derived in [25] is that the d-separation property or the *global directed Markov condition* of Definition IV.1 is recovered whenever all the random variables are discrete and the structural equations of the causal model satisfy a property known as *ancestrally unique solvability* (anSEP). Roughly, this property demands that the structural equation for each node must admit a unique solution given the values of the node's ancestors. We need not define this concept formally for our purposes here.

Ideally we would like to extend these ideas to quantum and post-quantum cyclic causal structures, where the causal mechanisms involve measurements and transformations on non-classical systems, which cannot be expressed using deterministic structural equations. In the non-classical case, it is unclear what conditions allow for the d-separation condition to be recovered. Even to make this question precise in the non-classical case, one would need to specify the analog of structural equations for such causal models which is an open problem. Here, we present a possible method for achieving this by explaining it using the following example and sketching how it might generalise to a larger class of causal models.

Example C.1 (A quantum cyclic causal model)**.** Consider the cyclic variation of the bipartite Bell causal structure illustrated in Figure 15a. Let the common cause Λ correspond to the Bell state $|\psi_\Lambda\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$. Suppose that A and B are the settings of local measurements on the two subsystems such that when these variables take the value 0, it denotes a measurement in the $\{|0\rangle, |1\rangle\}$ basis on the associated subsystem, and the value 1 denotes a measurement in the $\{|+\rangle, |-\rangle\}$ basis. X and Y are the binary outcomes of these measurements where $|0\rangle$ or $|+\rangle$ correspond to outcome 0 and $|1\rangle$ or $|-\rangle$ correspond to 1. The additional constraints coming from the causal loop are that $B = X$ and $A = Y$. This specifies all the causal mechanisms, how do we calculate the observed distribution P_{XYAB}?

a. A method based on post-selection: One method is to first calculate the observed correlations for the specified state and measurements in the original Bell scenario (Figure 1b), and then post-select on the observations that obey the loop conditions $B = X$ and $A = Y$. More formally, this corresponds to transforming the original cyclic causal structure of Figure 15a to the acyclic causal structure of Figure 15b by cutting off the edges $A \rightsquigarrow X$ and $B \rightsquigarrow Y$ and replacing them with the edges $A^* \rightsquigarrow X$ and $B^* \rightsquigarrow Y$ by introducing two exogenous nodes A^* and B^*. Then the inputs A^* and B^* and outputs X and Y along with the shared system Λ define a Bell scenario, while the variables $A = Y$ and $B = X$ can simply be seen as local post-processings of the outcomes. We can then calculate the observed probabilities for this acyclic causal structure using the Born rule, and post-select on $A^* = A$ and $B^* = B$, which effectively achieves the post-selection $A = Y$ and $B = X$ in the original Bell scenario (Figure 1b). This distribution needs to be renormalised to obtain the observed distribution P_{XYAB}. This is calculated in Figure 15 and can be used to find all the affects relations. An intervention on A would cut off the arrow from Y to A. This means A does not affect X since A is effectively exogenous in the post-intervention causal structure and will be uncorrelated with

X since Λ is the maximally entangled state. Similarly B does not affect Y. However, AB affects XY since a joint intervention on A and B takes us back to the original Bell scenario in which these sets are correlated, and correlation in the post-intervention causal structure implies an affects relation (cf. Lemma IV.2) and it can be checked that this affects relation is irreducible. We also have X affects B and Y affects A due to the loop conditions $A = Y$ and $B = X$. In addition, XY affects AB, which is also irreducible. With a bit more effort, one can also check that we have A affects Y and B affects X. Therefore we have two Type 1 affects causal loops (Definition VI.3) $\{A$ affects Y, Y affects $A\}$, and similarly for B and X. We also have a Type 4 affects causal loop (Definition VI.6) formed by the irreducible relations $\{AB$ affects XY, XY affects $AB\}$. In this example, $\mathcal{G}_{\mathrm{do}(A,B)}$ corresponds to a quantum causal structure (the Bell scenario) while $\mathcal{G}_{\mathrm{do}(X,Y)}$ is a simple classical causal structure. Then the (observed) arrows $\rightsquigarrow$ of $\mathcal{G}$ can be classified into dashed and solid arrows as: $A \dashrightarrow X$, $B \dashrightarrow Y$, $X \longrightarrow B$ and $Y \longrightarrow A$. The post-intervention distribution is fully specified here because, all interventions (except that on Λ alone) are associated with acyclic post-intervention graphs and for interventions on the exogenous Λ, the post and pre-intervention distributions coincide.

b. Applying the method to fine-tuned explanations of non-classical correlations: It is known that certain non-classical correlations arising in the bipartite Bell causal structure cannot be obtained in the same causal structure if the common cause Λ was classical. However, these correlations can be easily generated in the classical, fine-tuned causal structure of Figure 15c, which differs from the original causal structure by the inclusion of fine-tuned causal influences from each party's input to the other party's output. We now explain how this is achieved and then apply the post-selection method explained above to create a causal loop in Figure 15c by adding $X \longrightarrow B$ and $Y \longrightarrow A$. This will demonstrate that, even though the same non-classical correlations and affects relations can be obtained in the original Bell causal structure and its fine-tuned classical counterpart 15c, the two causal structures behave differently in the presence of causal loops.

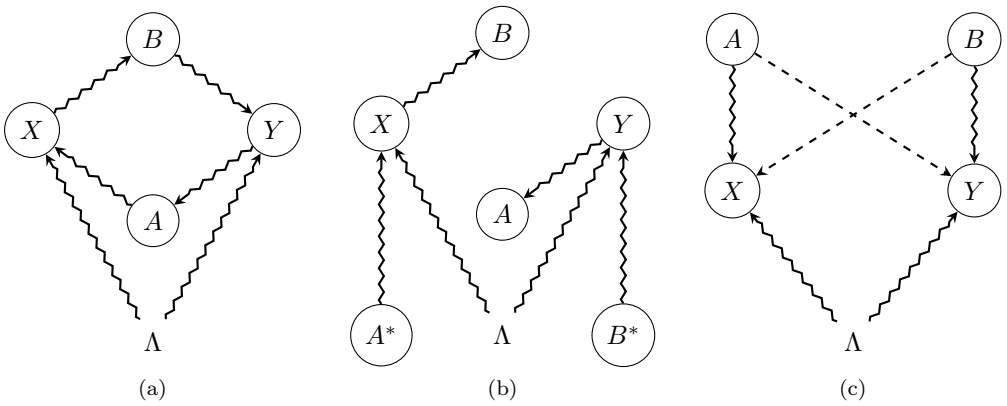
First consider the PR box, which is one of the maximally non-classical correlations of the Bell causal structure. It is defined by the condition $X \oplus Y = A.B$ where all the variables are binary. This is easily generated in the classical causal structure of Figure 15c by the structural equations $\Lambda = E$, $Y = E$ and $X = E \oplus A.B$ (where E is binary and uniformly distributed). Other non-classical correlations can be obtained by adding some "noise" to this PR box example. Let $\Lambda = (E, F)$ correspond to two variables E and F both binary, and the former distributed uniformly. Then the structural equations $Y = E$ and $X = E \oplus F \oplus A.B$ for different distributions over the exogenous variable F correspond to the PR box mixed with different levels of noise.[23]

$$\begin{aligned} X &= A.B \oplus E \oplus F, \\ Y &= E. \end{aligned} \tag{C1}$$

Therefore, the causal mechanisms that allow us to produce non-classical correlations P_{XYAB} in the acyclic causal structure 15c are the functional dependences (C1) along with a specification of the distributions over the exogenous variables E and F that constitute Λ. E is uniform while F can vary depending on the correlation to be generated. We now construct the causal loop by including the additional arrows $X \rightsquigarrow B$ and $Y \rightsquigarrow A$ and by effectively post-selecting on the loop condition $A = Y$ and $B = X$. These, along with the causal mechanisms (C1) of the acyclic case define the mechanisms for the cyclic causal structure. We will now see that these causal mechanisms are incompatible with each other. We have $Y = E$, $X = E \oplus F \oplus A.B$, $A = Y$ and $B = X$, which gives $X = E.X \oplus E \oplus F$ and $Y = E$. Therefore for $(E, F) = (0, 0)$, we have $(X, Y) = (0, 0)$ and for $(E, F) = (0, 1)$ we have $(X, Y) = (1, 0)$. However for $(E, F) = (1, 0)$ we get $X = X \oplus 1$ which does not have a solution. For $(E, F) = (1, 1)$ we get $X = X$ which is not a unique solution. Therefore if we demand unique solvability, we must require $E = 0$ deterministically which contradicts the initial assumption that E is uniform. Even if we do not require uniqueness, we can not have $(E, F) = (1, 0)$ and forbidding this would make E and F correlated and non-uniform.

Therefore, in the classical, fine-tuned explanation of the Bell correlations, adding the loop is not consistent with the causal mechanisms that generate the non-classical correlations in the absence of the loop— in particular, they are in conflict with the preparation of the exogenous variable Λ. If we have a consistent loop, then intervention on A and B will no longer recover the original non-classical correlations. This is in contrast to the faithful case analysed in Figure 15 (and explained previously in the text), when $\mathrm{do}(A, B)$ gives back the non-classical correlations of the Bell scenario. This suggests that certain (non-local) hidden variable explanations for quantum correlations (in a Bell experiment) can in principle be distinguished from the explanation provided by standard quantum mechanics in the presence of causal loops. We have only shown this for a particular set of functions or causal mechanisms for generating the former and it would be interesting to consider if this generalises, in particular to causal mechanisms provided by Bohmian mechanisms [37], a non-local hidden variable theory.

[23] Note that the model can be symmetrised by including an additional, uniformly distributed binary variable G in the description of $\Lambda = (E, F, G)$ and using the structural equations $X = E \oplus (G \oplus 1)(A.B \oplus F)$ and $Y = E \oplus G(A.B \oplus F)$.



(a) (b) (c)

X	Y	A	B	Measurements, outcomes	P^{QM}_{XYAB}	P_{XYAB}
0	0	0	0	$\sigma_Z \otimes \sigma_Z$, $(0,0)$	$\frac{1}{8}$	$\frac{1}{3}$
0	1	1	0	$\sigma_X \otimes \sigma_Z$, $(+,1)$	$\frac{1}{16}$	$\frac{1}{6}$
1	0	0	1	$\sigma_Z \otimes \sigma_X$, $(1,+)$	$\frac{1}{16}$	$\frac{1}{6}$
1	1	1	1	$\sigma_X \otimes \sigma_X$, $(-,-)$	$\frac{1}{8}$	$\frac{1}{3}$

(d)

FIG. 15: **A cyclic quantum causal model: (a)** A cyclic variation of the bipartite Bell causal structure (Figure 1b). **(b)** A method to calculate the observed distribution of (a) when Λ is non-classical involves this intermediate causal structure. This is obtained from (a) by copying the nodes A and B and removing the directed cycle as shown. This gives an acyclic causal structure for which the distribution $P_{XYABA^*B^*}$ can be calculated using known methods. Then, post-selecting on $A = A^*$ and $B = B^*$ gives the distribution P_{XYAB} for the original cyclic causal structure of (a). **(c)** A classical causal, fine-tuned structure that can generate, all non-classical correlations of the bipartite Bell causal structure. Creating a causal loop in this case by adding the arrows $X \rightsquigarrow B$ and $Y \rightsquigarrow A$ does not lead to the same predictions as (a), which corresponds to adding these arrows to the original Bell causal structure. This method explained in the main text. **(d)** The table provides the observed distribution for Example C.1 calculated using the proposed method. The only values of A, B, X and Y that are compatible with the loop conditions $A = Y$ and $B = X$ are those listed here, and the fifth column lists the measurements and outcomes that these values correspond to, according to Example C.1. P^{QM}_{XYAB} denotes the probabilities of the measurements and outcomes listed in the fifth column calculated using the Born rule. These values are sub-normalised, and upon renormalisation, the observed distribution P_{XYAB} for the cyclic causal structure (a) is obtained. Note that the d-separation condition IV.1 is satisfied in this case.

c. Generalising to other causal structures: The idea behind the post-selection method employed for Example C.1 above can in principle be generalised to other non-classical, cyclic causal structures where every directed cycle includes at least one edge $W \rightsquigarrow Z$ connecting classical nodes W and Z. The intuition is that cutting off such an edge in every directed cycle and replacing it with an edge $W^* \rightsquigarrow Z$, by introducing an additional, exogenous variable W^* would result in a directed acyclic graph (DAG). One can then apply the generalised causal model framework of [17] to obtain the observed distribution in this DAG and then post-select on $W = W^*$ for all the edges that were cut off. Then a way to recover the d-separation condition (using the result of [25]) would be to check whether these exists a classical causal model for the same cyclic causal structure that produces identical observed correlations and satisfies the anSEP property. Note that this classical causal model need not yield the same post-intervention distributions. In the example of Figure 15a, an intervention on A and B gives the Bell scenario, which as we know produces non-classical correlations that cannot be obtained in the corresponding classical causal model [11]. Finally, it would be interesting to compare this method with the framework of post-selected closed time-like curves [63].

Remark C.1. We note that assuming the d-separation condition of Definition IV.1 as a primitive property of the framework rules out certain cyclic causal structures from being described in our current framework. In the classical case, these are precisely those cyclic causal models that do not satisfy anSEP or those involving continuous random variables (due to the result of [25]). An example of such a causal model is given in [78], and [25] proposes a generalisation of d-separation called σ−separation through which they derive a *generalised global directed Markov* condition that applies to classical causal models involving continuous variables and/or do not satisfy anSEP. This reduces to d-separation in the acyclic case. Therefore, one option would be to replace d-separation with σ-separation in Definition IV.1 to generalise our framework for cyclic causal models. Doing so would not affect the results of the main paper, but would only enlarge the class of causal models to which they can be applied.

Appendix D: Proofs of all results

1. Proofs of Lemma IV.1 and Theorem IV.1

Lemma IV.1. *Let S_1, S_2 and S_3 be three disjoint sets of RVs such that $S_1 \perp\!\!\!\perp S_2|S_3$. If S is a set of RVs that is d-separated from these sets in a directed graph $\mathcal{G}$ containing all the members of S_1, S_2, S_3 and S as nodes i.e., $S \perp^d S_i$ $\forall i \in \{1,2,3\}$, then any distribution P that is compatible with $\mathcal{G}$ also satisfies the following conditional independences,*

$$S_1 S \perp\!\!\!\perp S_2|S_3, \quad S_1 \perp\!\!\!\perp S_2 S|S_3 \text{ and } \quad S_1 \perp\!\!\!\perp S_2|S_3 S.$$

Proof. The conditional independence $S_1 \perp\!\!\!\perp S_2|S_3$ stands for $P_{S_1S_2|S_3} = P_{S_1|S_3}P_{S_2|S_3}$, which implies

$$P_{S_1|S_2S_3} = P_{S_1|S_3}. \tag{D1}$$

The three d-separation relations $S \perp^d S_i$ for $i \in \{1,2,3\}$ imply that S is d-separated from every subset of the union $S_1S_2S_3$. This implies the following independences by Definition IV.1 of compatibility of the distribution P with the causal model represented by $\mathcal{G}$,

$$P_{S|S'} = P_S \quad \forall S' \subseteq S_1S_2S_3. \tag{D2}$$

Now consider the conditional distribution $P_{S_2|SS_1S_3}$. We have,

$$\begin{aligned} P_{S_2|SS_1S_3} &= \frac{P_{S_2SS_1S_3}}{P_{SS_1S_3}} \\ &= \frac{P_{S_3}P_{S_2|S_3}P_{S_1|S_2S_3}P_{S|S_1S_2S_3}}{P_{SS_1S_3}} \\ &= \frac{P_{S_3}P_{S_2|S_3}P_{S_1|S_3}P_S}{P_SP_{S_1S_3}} \\ &= P_{S_2|S_3}, \end{aligned} \tag{D3}$$

where we have used Equations (D1) and (D2) in the third line, noting that $P_{S|S_1S_3} = P_S \Rightarrow P_{SS_1S_3} = P_SP_{S_1S_3}$. Equation (D3) is equivalent to $P_{SS_1S_2|S_3} = P_{SS_1|S_3}P_{S_2|S_3}$ which denotes the conditional independence $SS_1 \perp\!\!\!\perp S_2|S_3$. The conditional independence $S_1 \perp\!\!\!\perp SS_2|S_3$ can be derived analogously due to the symmetry between S_1 and S_2.

Finally, we have

$$P_{S_2|SS_3} = \frac{P_{S_2SS_3}}{P_{SS_3}} = \frac{P_{S_3}P_{S_2|S_3}P_{S|S_2S_3}}{P_SP_{S_3}} = P_{S_2|S_3}, \tag{D4}$$

and similarly $P_{S_1|SS_3} = P_{S_1|S_3}$. Together with Equation (D3), this implies $P_{S_2|SS_1S_3} = P_{S_2|SS_3}$. This is equivalent to $P_{S_1S_2|SS_3} = P_{S_1|SS_3}P_{S_2|SS_3}$ which denotes the final conditional independence $S_1 \perp\!\!\!\perp S_2|SS_3$. □

Theorem IV.1. *Given a causal model over a set S of observed nodes, associated causal graph $\mathcal{G}$ and a distribution P_S compatible with $\mathcal{G}$ according to Definition IV.1, the following 3 rules of do-calculus [2] hold for interventions on this causal model.*

- ***Rule 1: Ignoring observations***

$$P_{\mathcal{G}_{\text{do}(X)}}(y|x,z,w) = P_{\mathcal{G}_{\text{do}(X)}}(y|x,w) \qquad \text{if } (Y \perp^d Z|XW)_{\mathcal{G}_{\overline{X}}} \tag{5}$$

- ***Rule 2: Action/observation exchange***

$$P_{\mathcal{G}_{\mathrm{do}(XZ)}}(y|x,z,w) = P_{\mathcal{G}_{\mathrm{do}(X)}}(y|x,z,w) \qquad \text{if } (Y \perp^d Z|XW)_{\mathcal{G}_{\overline{X}\underline{Z}}} \tag{6}$$

- ***Rule 3: Ignoring actions/interventions***

$$P_{\mathcal{G}_{\mathrm{do}(XZ)}}(y|x,z,w) = P_{\mathcal{G}_{\mathrm{do}(X)}}(y|x,w) \qquad \text{if } (Y \perp^d Z|XW)_{\mathcal{G}_{\overline{XZ(W)}}}, \tag{7}$$

where X, Y, Z and W are disjoint subsets of the observed nodes, $Z(W)$ denotes the set of nodes in Z which are not ancestors of W, and the above hold for all values x, y, z and w of X, Y, Z and W.

Proof. **Rule 1:** We first note that the graph $\mathcal{G}_{\mathrm{do}(X)}$ differs from $\mathcal{G}_{\overline{X}}$ only by the inclusion of the additional nodes I_{X_i} and corresponding edge $I_{X_i} \rightsquigarrow X_i$ for each $X_i \in X$. Therefore, the d-separation relation $(Y \perp^d Z|XW)_{\mathcal{G}_{\overline{X}}}$ for the latter implies the same relation $(Y \perp^d Z|XW)_{\mathcal{G}_{\mathrm{do}(X)}}$ for the former. Using the compatibility condition of Definition IV.1 for the graph $\mathcal{G}_{\mathrm{do}(X)}$, this implies the conditional independence of Y and Z given XW for the distribution $P_{\mathcal{G}_{\mathrm{do}(X)}}$ i.e., $P_{\mathcal{G}_{\mathrm{do}(X)}}(y,z|x,w) = P_{\mathcal{G}_{\mathrm{do}(X)}}(y|x,w)P_{\mathcal{G}_{\mathrm{do}(X)}}(z|x,w)$. This conditional independence is equivalently expressed by the required Equation (5).

Rule 2: $\mathcal{G}_{\overline{X},\underline{Z}}$ is the graph where all incoming arrows to X and outgoing arrows from Z are removed in $\mathcal{G}$. Hence, the d-separation condition $(Y \perp^d Z|XW)_{\mathcal{G}_{\overline{X},\underline{Z}}}$ implies that the only paths between Y and Z in the graph $\mathcal{G}_{\overline{X}}$ that are not blocked by X and W are paths involving an outgoing arrow from Z. These are precisely the paths that get removed in going from $\mathcal{G}_{\overline{X}}$ to $\mathcal{G}_{\overline{X},\underline{Z}}$, resulting in the required d-separation there. The same statement holds for the graph $\mathcal{G}_{\mathrm{do}(X)}$ (by the argument used in the proof of Rule 1), and also for the graph $\mathcal{G}_{\mathrm{do}(X),I_Z}$ which corresponds to adding the nodes I_{Z_i} and edges $I_{Z_i} \rightsquigarrow Z_i$ to $\mathcal{G}_{\mathrm{do}(X)}$ for each $Z_i \in Z$. The latter holds true since the addition of the I_{Z_i} nodes and $I_{Z_i} \rightsquigarrow Z_i$ edges cannot create any additional paths between Z and Y that are left unblocked by X and W. This implies that the only paths between Y and the set $I_Z := \{I_{Z_i}\}_i$ not blocked by X and W in $\mathcal{G}_{\mathrm{do}(X),I_Z}$ are paths from I_Z, going through Z and involving an outgoing arrow from Z i.e., paths involving the subgraph $I_Z \rightsquigarrow Z \rightsquigarrow \dots$. All these paths would get blocked when conditioning additionally on Z. This gives $(Y \perp^d I_Z|XWZ)_{\mathcal{G}_{\mathrm{do}(X),I_Z}}$, which through the compatibility condition (Definition IV.1) implies the conditional independence $(Y \perp\!\!\!\perp I_Z|XWZ)_{\mathcal{G}_{\mathrm{do}(X),I_Z}}$, equivalently expressed as

$$P_{\mathcal{G}_{\mathrm{do}(X),I_Z}}(y|x,w,z,I_Z = \mathrm{idle}) = P_{\mathcal{G}_{\mathrm{do}(X),I_Z}}(y|x,w,z,I_Z = \mathrm{do}(z)) \quad \forall y,x,w,z. \tag{D5}$$

Using Equations 4a and 4b, we have $P_{\mathcal{G}_{\mathrm{do}(X),I_Z}}(y|x,w,z,I_Z = \mathrm{idle}) = P_{\mathcal{G}_{\mathrm{do}(X)}}(y|x,w,z)$ and $P_{\mathcal{G}_{\mathrm{do}(X),I_Z}}(y|x,w,z,I_Z = \mathrm{do}(z)) = P_{\mathcal{G}_{\mathrm{do}(XZ)}}(y|x,w,z)$ respectively $\forall y,x,w,z$. Along with Equation (D5), this gives the required Equation 6. In other words, once X, W and Z are given, Y does not depend on whether the given value z of Z was obtained through an intervention ($I_Z = \mathrm{do}(z)$) or passive observation (i.e., where $I_{Z_i} = \mathrm{idle}$ for all i, which is the causal model where no interventions are made on elements of Z).

Rule 3: Consider the graph $\mathcal{G}_{\mathrm{do}(X)I_Z}$ which is the post-intervention graph with respect to the nodes X augmented with $I_{Z_i} \rightsquigarrow Z_i$ for all $Z_i \in Z$. In this graph, suppose we had the d-separation relation $(Y \perp^d I_Z|XW)_{\mathcal{G}_{\mathrm{do}(X)I_Z}}$. By Definition IV.1, this would result in the conditional independence $(Y \perp\!\!\!\perp I_Z|XW)_{\mathcal{G}_{\mathrm{do}(X)I_Z}}$ which can be expressed as

$$P_{\mathcal{G}_{\mathrm{do}(X)I_Z}}(Y|W,X,I_Z = \mathrm{idle}) = P_{\mathcal{G}_{\mathrm{do}(X)I_Z}}(Y|W,X,I_Z = \mathrm{do}(z)) \quad \forall z$$

Using the defining rules (4a) and (4b) then gives $P_{\mathcal{G}_{\mathrm{do}(X)I_Z}}(Y|W,X,I_Z = \mathrm{idle}) = P_{\mathcal{G}_{\mathrm{do}(X)}}(Y|W,X)$ and $P_{\mathcal{G}_{\mathrm{do}(X)I_Z}}(Y|W,X,I_Z = \mathrm{do}(z)) = P_{\mathcal{G}_{\mathrm{do}(XZ)}}(Y|W,X,Z = z)$ $\forall z$, and consequently $P_{\mathcal{G}_{\mathrm{do}(XZ)}}(Y|W,X,Z) = P_{\mathcal{G}_{\mathrm{do}(X)}}(Y|W,X)$ which is the required Equation (7). Therefore, showing that the d-separation condition $(Y \perp^d Z|XW)_{\mathcal{G}_{\overline{X},\overline{Z(W)}}}$ implies the d-separation relation $(Y \perp^d I_Z|XW)_{\mathcal{G}_{\mathrm{do}(X)I_Z}}$ would complete the proof. This is shown by contradiction. Suppose that $(Y \perp^d Z|XW)_{\mathcal{G}_{\overline{X},\overline{Z(W)}}}$ and $(Y \not\perp^d I_Z|XW)_{\mathcal{G}_{\mathrm{do}(X)I_Z}}$. Then there must exist a path from a member I_{Z_i} of I_Z to a member Y_j of Y in $\mathcal{G}_{\mathrm{do}(X)I_Z}$ that is unblocked by X and W. There are two possibilities for such a path: either it contains the subgraph $I_{Z_i} \rightsquigarrow Z_i \rightsquigarrow \dots Y_j$ or the subgraph $I_{Z_i} \rightsquigarrow Z_i \leftsquigarrow \dots Y_j$. Denoting these possibilities as cases 1 and 2 respectively, let $\mathscr{P}$ be the shortest such path. We will show that a contraction arises in each case.

Case 1: Consider the first case where $\mathscr{P}$ contains the subgraph $I_{Z_i} \rightsquigarrow Z_i \rightsquigarrow \dots Y_j$. Note $(Y \not\perp^d I_Z | XW)_{\mathcal{G}_{\text{do}(X)I_Z}}$ (which we have assumed) implies $(Y \not\perp^d Z_i | XW)_{\mathcal{G}_{\text{do}(X)}}$. Along with the assumption that $(Y \perp^d Z_i | XW)_{\mathcal{G}_{\overline{X},\overline{Z(W)}}}$, this implies that there exists a path from Z_i to Y in $\mathcal{G}_{\text{do}(X)}$ unblocked by X and W that passes through some member Z_k of $Z(W)$ which would blocked when the incoming arrows to Z_k are removed. This leads to the following subcases where the path from Z_i to Y_j in $\mathcal{G}_{\text{do}(X)}$ contains the following subgraphs:

- *Case 1a:* $Z_i \rightsquigarrow \dots \rightsquigarrow Z_k \leftsquigarrow \dots Y_j$ or
- *Case 1b:* $Z_i \rightsquigarrow \dots \leftsquigarrow Z_k \leftsquigarrow \dots Y_j$ or
- *Case 1c:* $Z_i \rightsquigarrow \dots \rightsquigarrow Z_k \rightsquigarrow \dots Y_j$

None of these can occur for the following reasons. In *Case 1a*, some descendant of Z_k must be in W for the path to be unblocked in $\mathcal{G}_{\text{do}(X)}$ but by definition, $Z(W)$ (which contains Z_k) is the set of all nodes in Z that do not have descendants in W. In *Case 1b*, the path between Z_i and Z_k must contain a collider. For this path to be unblocked by X and W in $\mathcal{G}_{\text{do}(X)}$ the collider node must have a descendant in W but the other requirement that this path must be blocked in $\mathcal{G}_{\overline{X},\overline{Z(W)}}$ implies that the same collider node must be a member of $Z(W)$ which by definition does not have any descendants in W, yielding a contradiction. In *Case 1c*, there is either a directed path from $Z_k \in Z$ to $Y_j \in Y$ in $\mathcal{G}_{\text{do}(X)}$ or a collider in the path between Z_k and Y_j. The latter is ruled out by the same argument used in Case 1b. If there is a directed path from Z_k to Y_j, then there is a directed path from I_{Z_k} to Y_j in $\mathcal{G}_{\text{do}(X)I_Z}$ i.e., there is a path from a member of Z to Y that is unblocked by X and W in $\mathcal{G}_{\text{do}(X)I_Z}$ and that is shorter than the shortest path $\mathscr{P}$, which is not possible.

Finally, consider *Case 2* where the path $\mathscr{P}$ contains the subgraph $I_{Z_i} \rightsquigarrow Z_i \leftsquigarrow \dots Y_j$. The initial assumption that $(Y \not\perp^d I_Z | XW)_{\mathcal{G}_{\text{do}(X)I_Z}}$ implies that the collider node Z_i must have descendants in the conditioning set W i.e., $Z_i \notin Z(W)$. However, in this case we will violate the assumption that $(Y \perp^d Z | XW)_{\mathcal{G}_{\overline{X},\overline{Z(W)}}}$. On the other hand, to satisfy this d-separation, we would require $Z_i \in Z(W)$ but this would violate $(Y \not\perp^d I_Z | XW)_{\mathcal{G}_{\text{do}(X)I_Z}}$. Hence we have shown that $(Y \perp^d Z | XW)_{\mathcal{G}_{\overline{X},\overline{Z(W)}}}$ and $(Y \not\perp^d I_Z | XW)_{\mathcal{G}_{\text{do}(X)I_Z}}$ can never be simultaneously satisfied and hence that $(Y \perp^d Z | XW)_{\mathcal{G}_{\overline{X},\overline{Z(W)}}}$ implies $(Y \not\perp^d I_Z | XW)_{\mathcal{G}_{\text{do}(X)I_Z}}$ which in turn implies the required Equation (7).

□

2. Proofs of Lemmas IV.3, IV.4, IV.5, IV.6, IV.7, IV.8 and Corollary IV.3

Lemma IV.3. *For a causal model over a set S of RVs where X, Y, Z and W are any pairwise disjoint subsets of S,*

1. *X affects Y given $do(Z) \Rightarrow X$ is a cause of Y (cf. Definition II.1).*
2. *X affects Y given $\{\text{do}(Z), W\} \Rightarrow X$ is a cause of Y or X is a cause of W.*

Proof. 1. We prove this by contradiction. The relation X is not a cause of Y is equivalent to the absence of any directed paths from X to Y in $\mathcal{G}$ i.e., $(X \perp^d Y)_{\mathcal{G}_{\text{do}(X)}}$ and consequently $(X \perp^d Y)_{\mathcal{G}_{\text{do}(XZ)}}$, for any subset Z of observed nodes, pairwise disjoint to X and Y. Since Z is effectively exogenous in $\mathcal{G}_{\text{do}(XZ)}$, $(X \perp^d Y)_{\mathcal{G}_{\text{do}(XZ)}}$ implies $(X \perp^d Y | Z)_{\mathcal{G}_{\text{do}(XZ)}}$. Applying Rule 3 of Theorem IV.1 (noting the relation between $\mathcal{G}_{\bar{Z}\bar{X}}$ and $\mathcal{G}_{\text{do}(XZ)}$) to the latter implies that $P_{\mathcal{G}_{\text{do}(XZ)}}(Y|X,Z) = P_{\mathcal{G}_{\text{do}(Z)}}(Y|Z)$ which is equivalent to X does not affect Y given $\text{do}(Z)$.

2. This follows from the first part of Lemma IV.8 (proven later in this appendix) and the first part of this lemma. By Lemma IV.8, X affects Y given $\{\text{do}(Z), W\}$ implies X affects YW given $\text{do}(Z)$, which in turn implies that X must either be a cause of Y or of W, by the first part, proven above.

□

Lemma IV.4. *For a causal model over a set S of RVs where X, Y, Z and W are pairwise disjoint subsets of S,*

$$X \text{ affects } Y \text{ given } \{\text{do}(Z), W\} \quad \Rightarrow \quad Z \text{ affects } Y \text{ given } W \quad \text{or} \quad XZ \text{ affects } Y \text{ given } W.$$

Proof. To establish the lemma, we show that it is not possible to have Z does not affect Y given W *and* XZ does not affect Y given W whenever X affects Y given $\{\text{do}(Z), W\}$. Writing these three conditions out we have

$$P_{\mathcal{G}_{\text{do}(Z)}}(Y|Z,W) = P_{\mathcal{G}}(Y|W), \tag{D6a}$$

$$P_{\mathcal{G}_{\mathrm{do}(XZ)}}(Y|X,Z,W) = P_{\mathcal{G}}(Y|W), \tag{D6b}$$

$$P_{\mathcal{G}_{\mathrm{do}(XZ)}}(Y|X,Z,W) \neq P_{\mathcal{G}_{\mathrm{do}(Z)}}(Y|Z,W). \tag{D6c}$$

Equations (D6a) and (D6b) imply $P_{\mathcal{G}_{\mathrm{do}(XZ)}}(Y|X,Z,W) = P_{\mathcal{G}_{\mathrm{do}(Z)}}(Y|Z,W)$ in contradiction with Equation (D6c). □

Lemma IV.5. *For a causal model over a set S of RVs where X, Y, Z and W are pairwise disjoint subsets of S and X consists only of exogenous nodes,*

$$X \text{ affects } Y \text{ given } \{\mathrm{do}(Z), W\} \quad \Rightarrow \quad XZ \text{ affects } Y \text{ given } W.$$

Proof. By Lemma IV.4, if X affects Y given $\{\mathrm{do}(Z), W\}$ then there are only three possibilities 1) Z affects Y given W and XZ does not affect Y given W, 2) Z does not affect Y given W and XZ affects Y given W, and 3) Z affects Y given W and XZ affects Y given W i.e., the only case where the required conclusion does not follow is 1). Then the proof will be complete if we show that whenever X consists only of exogenous nodes the undesired case does not arise. We show this by establishing that for exogenous X, Z affects Y given W implies XZ affects Y given W. Suppose by contradiction that XZ does not affect Y given W i.e., $P_{\mathcal{G}_{\mathrm{do}(XZ)}}(Y|X,Z,W) = P_{\mathcal{G}}(Y|W)$. By the exogeneity of X, this becomes $P_{\mathcal{G}_{\mathrm{do}(Z)}}(Y|X,Z,W) = P_{\mathcal{G}}(Y|W)$ or equivalently, $P_{\mathcal{G}_{\mathrm{do}(Z)}}(Y,X,Z,W) = P_{\mathcal{G}}(Y|W)P_{\mathcal{G}_{\mathrm{do}(Z)}}(X,Z,W)$. Summing over values of X and rearranging gives $P_{\mathcal{G}_{\mathrm{do}(Z)}}(Y|Z,W) = P_{\mathcal{G}}(Y|W)$ which is equivalent to Z does not affect Y given W. Therefore Z affects Y given W implies XZ affects Y given W whenever X is exogenous. □

Lemma IV.6. *For every reducible affects relation X affects Y given $\{\mathrm{do}(Z), W\}$, there exists a proper subset $\tilde{s}_X$ of X such that $\tilde{s}_X$ affects Y given $\{\mathrm{do}(Z), W\}$.*

Proof. By definition, if X affects Y given $\{\mathrm{do}(Z), W\}$ is reducible, then there exists a proper subset s_X of X such that s_X does not affect Y given $\{\mathrm{do}(Z\tilde{s}_X), W\}$. We now show that for every such s_X, its complement $\tilde{s_X} := X \backslash s_X$ is such that $\tilde{s}_X$ affects Y given $\{\mathrm{do}(Z), W\}$. We show this by contradiction. Assume that $\tilde{s}_X$ does not affect Y given $\{\mathrm{do}(Z), W\}$ while X affects Y given $\{\mathrm{do}(Z), W\}$ and s_X does not affect Y given $\{\mathrm{do}(Z\tilde{s}_X), W\}$. Explicitly, these correspond to the following conditions, noting that $s_X\tilde{s}_X = X$:

$$P_{\mathcal{G}_{\mathrm{do}(\tilde{s}_X Z)}}(Y|\tilde{s}_X,Z,W) = P_{\mathcal{G}_{\mathrm{do}(Z)}}(Y|Z,W), \tag{D7a}$$

$$P_{\mathcal{G}_{\mathrm{do}(XZ)}}(Y|X,Z,W) \neq P_{\mathcal{G}_{\mathrm{do}(Z)}}(Y|Z,W), \tag{D7b}$$

$$P_{\mathcal{G}_{\mathrm{do}(XZ)}}(Y|X,Z,W) = P_{\mathcal{G}_{\mathrm{do}(\tilde{s}_X Z)}}(Y|\tilde{s}_X,Z,W). \tag{D7c}$$

Equations (D7a) and (D7c) imply that $P_{\mathcal{G}_{\mathrm{do}(XZ)}}(Y|X,Z,W) = P_{\mathcal{G}_{\mathrm{do}(Z)}}(Y|Z,W)$, which contradicts Equation (D7b). □

Lemma IV.7. *For a causal model over a set S of RVs of which X_1, X_2, Y, Z and W are pairwise disjoint subsets,*

$$X_1 \text{ affects } Y \text{ given } \{\mathrm{do}(Z), W\} \text{ and } X_2 \text{ does not affect } Y \text{ given } \{\mathrm{do}(ZX_1), W\}$$
$$\Downarrow$$
$$X_1X_2 \text{ affects } Y \text{ given } \{\mathrm{do}(Z), W\}.$$

Proof. The proof is similar to that of Lemma IV.6. X_1 affects Y given $\{\mathrm{do}(Z), W\}$ and X_2 does not affect Y given $\{\mathrm{do}(ZX_1), W\}$ are equivalent to

$$P_{\mathcal{G}_{\mathrm{do}(X_1 Z)}}(Y|X_1,Z,W) \neq P_{\mathcal{G}_{\mathrm{do}(Z)}}(Y|Z,W), \tag{D8a}$$

$$P_{\mathcal{G}_{\mathrm{do}(X_1X_2Z)}}(Y|X_1,X_2,Z,W) = P_{\mathcal{G}_{\mathrm{do}(X_1Z)}}(Y|X_1,Z,W). \tag{D8b}$$

These yield $P_{\mathcal{G}_{\mathrm{do}(X_1X_2Z)}}(Y|X_1,X_2,Z,W) \neq P_{\mathcal{G}_{\mathrm{do}(Z)}}(Y|Z,W)$ which is equivalent to X_1X_2 affects Y given $\{\mathrm{do}(Z), W\}$. □

Lemma IV.8. *For a causal model over a set S of RVs where X, Y, Z and W are pairwise disjoint subsets of S,*

1. *X affects Y given $\{\mathrm{do}(Z), W\} \Rightarrow X$ affects YW given $do(Z)$.*

2. *X affects Y given $\{\mathrm{do}(Z), W\}$ is irreducible $\Rightarrow$ X affects YW given $do(Z)$ is irreducible.*

3. *X affects YW given $do(Z)$ $\Leftrightarrow$ X affects Y given $\{\mathrm{do}(Z), W\}$ or X affects W given $do(Z)$.*

Proof. 1. We prove this through the contrapositive. Suppose that X does not affect YW given $\mathrm{do}(Z)$ i.e.,

$$P_{\mathcal{G}_{\mathrm{do}(XZ)}}(Y, W|X, Z) = P_{\mathcal{G}_{\mathrm{do}(Z)}}(Y, W|Z) \tag{D9}$$

Summing over values of Y on both sides, we have $P_{\mathcal{G}_{\mathrm{do}(XZ)}}(W|X, Z) = P_{\mathcal{G}_{\mathrm{do}(Z)}}(W|Z)$ i.e., X does not affect W given $\mathrm{do}(Z)$. Hence

$$\begin{aligned} P_{\mathcal{G}_{\mathrm{do}(XZ)}}(Y, W|X, Z)/P_{\mathcal{G}_{\mathrm{do}(XZ)}}(W|X, Z) &= P_{\mathcal{G}_{\mathrm{do}(Z)}}(Y, W|Z)/P_{\mathcal{G}_{\mathrm{do}(Z)}}(W|Z) \\ \Rightarrow P_{\mathcal{G}_{\mathrm{do}(XZ)}}(Y|X, Z, W) &= P_{\mathcal{G}_{\mathrm{do}(Z)}}(Y|Z, W), \end{aligned} \tag{D10}$$

which is equivalent to X does not affect Y given $\{\mathrm{do}(Z), W\}$.

2. Suppose that X affects Y given $\{\mathrm{do}(Z), W\}$ is irreducible i.e.,

$$P_{\mathcal{G}_{\mathrm{do}(XZ)}}(Y|X, Z, W) \neq P_{\mathcal{G}_{\mathrm{do}(\tilde{s}_X Z)}}(Y|\tilde{s}_X, Z, W), \qquad \forall s_X \subset X \tag{D11}$$

where $s_X \tilde{s}_X := X$. If X affects YW given $\mathrm{do}(Z)$ is reducible, then there exists a partition of $X = s_X \tilde{s}_X$ such that

$$P_{\mathcal{G}_{\mathrm{do}(XZ)}}(Y, W|X, Z) = P_{\mathcal{G}_{\mathrm{do}(\tilde{s}_X Z)}}(Y, W|\tilde{s}_X, Z). \tag{D12}$$

As in the proof of part 1., this implies that $P_{\mathcal{G}_{\mathrm{do}(XZ)}}(Y|X, Z, W) = P_{\mathcal{G}_{\mathrm{do}(\tilde{s}_X Z)}}(Y|\tilde{s}_X, Z, W)$, which contradicts the first equation. Therefore X affects Y given $\{\mathrm{do}(Z), W\}$ is irreducible implies X affects YW given $\mathrm{do}(Z)$ is irreducible.

3. For the forward direction, it is again convenient to use the contrapositive, i.e., to show that X does not affect Y given $\{\mathrm{do}(Z), W\}$ and X does not affect W given $\mathrm{do}(Z)$ imply X does not affect YW given $\mathrm{do}(Z)$. The first two statements are

$$\begin{aligned} P_{\mathcal{G}_{\mathrm{do}(XZ)}}(Y|X, Z, W) &= P_{\mathcal{G}_{\mathrm{do}(Z)}}(Y|Z, W) \quad \text{and} \\ P_{\mathcal{G}_{\mathrm{do}(XZ)}}(W|X, Z) &= P_{\mathcal{G}_{\mathrm{do}(Z)}}(W|Z). \end{aligned}$$

Multiplying these gives

$$P_{\mathcal{G}_{\mathrm{do}(XZ)}}(Y|X, Z, W) P_{\mathcal{G}_{\mathrm{do}(XZ)}}(W|X, Z) = P_{\mathcal{G}_{\mathrm{do}(Z)}}(Y|Z, W) P_{\mathcal{G}_{\mathrm{do}(Z)}}(W|Z)$$

which rearranges to

$$P_{\mathcal{G}_{\mathrm{do}(XZ)}}(Y, W|X, Z) = P_{\mathcal{G}_{\mathrm{do}(Z)}}(Y, W|Z),$$

which is X does not affect YW given $\mathrm{do}(Z)$.

For the reverse direction, we note that we have shown X affects W given $\mathrm{do}(Z)$ implies X affects YW given $\mathrm{do}(Z)$ in the proof of part 1 of this lemma. From the main statement of part 1. we also have X affects Y given $\{\mathrm{do}(Z), W\}$ implies X affects YW given $\mathrm{do}(Z)$. Therefore we have X affects Y given $\{\mathrm{do}(Z), W\}$ or X affects W given $\mathrm{do}(Z)$ implies X affects YW given $\mathrm{do}(Z)$. □

Corollary IV.3. *For a causal model over a set S of RVs where X, Y, Z and W are any pairwise disjoint subsets of S,*

1. *X affects Y given $do(Z)$ is irreducible $\Rightarrow$ for each element $e_X \in X$ there exists an element $e_Y \in Y$ such that e_X is a cause of e_Y.*

2. *X affects Y given $\{\mathrm{do}(Z), W\}$ is irreducible $\Rightarrow$ for each element $e_X \in X$ there exists an element $e_{YW} \in YW$ such that e_X is a cause of e_{YW}.*

Proof. 1. Given that X affects Y given $\mathrm{do}(Z)$ is irreducible, we know that for every $s_X \subset X$, s_X affects Y given $\mathrm{do}(Z\tilde{s}_X)$, where $s_X \tilde{s}_X := X$. In particular, this means that for every element $e_X \in X$, e_X affects Y given $\mathrm{do}(Z\tilde{e}_X)$. Then by using Lemma IV.3, we know that e_X is a cause of Y, which by Definition II.1 means that there exists a directed path from e_X to at least one element $e_Y \in Y$ which in turn means that e_X is a cause of e_Y.

2. By parts 1 and 2 of Lemma IV.8, X affects Y given $\{\mathrm{do}(Z), W\}$ implies X affects YW given $\mathrm{do}(Z)$ and the irreducibility of the former implies the irreducibility of the latter which in turn implies (by the first part of the current lemma) that for every $e_X \in X$, there exists $e_{YW} \in YW$ such that e_X is a cause of e_{YW}.

□

3. Proofs of Theorems V.1, VI.1 and Lemma VI.1

Theorem V.1. *[Necessary and sufficient conditions for compatibility with an embedding in $\mathcal{T}$] Let $\mathcal{S}$ be set of ORVs embedded in a partial order $\mathcal{T}$ with respect to an embedding $\mathscr{E}$ and let $\mathscr{A}$ be a given set of affects relations on $\mathcal{S}$. Further, consider forming an augmented set of ORVs $\mathcal{S}'$ by taking $\mathcal{S}$ and for each variable $\mathcal{X} \in \mathcal{S}$, embedding a copy of $\mathcal{X}$ at each point in its accessible region $\mathcal{R}_{\mathcal{X}}$ and form $\mathscr{A}'$ by adding to $\mathscr{A}$ that each variable affects each of its copies for all copies. Then the following statements hold.*

1. *If the set of affects relations $\mathscr{A}$ is compatible with the embedding $\mathscr{E}$ in $\mathcal{T}$, then* $\mathbf{compat1'}(\mathcal{S}', \mathscr{A}')$ *holds i.e.,* $\mathbf{compat1'}(\mathcal{S}', \mathscr{A}')$ *is necessary for compatibility of $\mathscr{A}$ with the space-time embedding $\mathscr{E}$.*
2. $\mathbf{compat1'}(\mathcal{S}', \mathscr{A}')$ *implies that $\mathcal{R}_{\mathcal{X}} \subseteq \overline{\mathcal{F}}(\mathcal{X})\ \forall \mathcal{X} \in \mathcal{S}$, but not that the two sets $\mathcal{R}_{\mathcal{X}}$ and $\overline{\mathcal{F}}(\mathcal{X})$ are necessarily equal $\forall \mathcal{X} \in \mathcal{S}$, i.e.,* $\mathbf{compat1'}(\mathcal{S}', \mathscr{A}')$ *is not sufficient for compatibility of $\mathscr{A}$ with the space-time embedding $\mathscr{E}$.*

Proof. 1. If $\mathbf{compat}(\mathcal{S}, \mathscr{A})$ holds then $\mathcal{R}_{\mathcal{X}} = \overline{\mathcal{F}}(\mathcal{X})$ for all $\mathcal{X} \in \mathcal{S}$. Hence by Definition V.5 of accessible regions for sets of ORVs, we have $\mathbf{compat1'}(\mathcal{S}, \mathscr{A})$. The remaining affects relations in $\mathscr{A}'$ are all of the form $\mathcal{X}$ affects $\mathcal{X}'$ where $\mathcal{X}'$ is a copy of $\mathcal{X}$, and so, since the location of $\mathcal{X}'$ is in $\mathcal{R}_{\mathcal{X}} = \overline{\mathcal{F}}(\mathcal{X})$, $\mathbf{compat1'}(\mathcal{S}', \mathscr{A}')$ also holds.

2. $\mathbf{compat1'}(\mathcal{S}', \mathscr{A}')$ when applied to the affects relations of the form $\mathcal{X}$ affects $\mathcal{X}'$ when $\mathcal{X}'$ is a copy of $\mathcal{X}$ tells us that $\overline{\mathcal{F}}(\mathcal{X}') \subseteq \overline{\mathcal{F}}(\mathcal{X})$ for every copy $\mathcal{X}'$ of $\mathcal{X}$, while Definition V.4 tells us that $\overline{\mathcal{F}}(\mathcal{X}') \subseteq \mathcal{R}_{\mathcal{X}}$ for every copy $\mathcal{X}'$ of $\mathcal{X}$. If $\mathcal{R}_{\mathcal{X}} \nsubseteq \overline{\mathcal{F}}(\mathcal{X})$ then it would be possible for a copy of X to be accessible outside its future, and hence that $\overline{\mathcal{F}}(\mathcal{X}') \nsubseteq \overline{\mathcal{F}}(\mathcal{X})$, contradicting $\mathbf{compat1'}(\mathcal{S}', \mathscr{A}')$. Therefore $\mathcal{R}_{\mathcal{X}} \subseteq \overline{\mathcal{F}}(\mathcal{X})$ must hold.

□

Theorem VI.1. *Any set of affects relations $\mathscr{A}$ containing an affects causal loop of Type 1, 2, 3, 4, 5 or 6 can only arise from a causal model over a cyclic causal structure i.e., these are indeed instances of affects causal loops according to Definition VI.1.*

Proof. Noting that all affects causal loops of Types 1, 2, 3 and 4 are also affects causal loops of Type 5, proving the theorem for ACL5 and ACL6 would imply the required result for ACL1, ..., ACL6.

1. *Proof for ACL5* Applying Corollary IV.3 to all affects relations in $S_i \subseteq \hat{S}_i$, $i = 1, \ldots, n$ such that $\{\hat{S}_1$ affects S_2, $\hat{S}_2$ affects S_3, ...,$\hat{S}_{n-1}$ affects S_n, $\hat{S}_n$ affects $S_1\} \subseteq \mathscr{A}$, we know that each element of $\hat{S}_i$ must be a cause of some element of $S_{i+1 \bmod n}$. Following the chain, this implies that each element $e^1 \in S_1 \subseteq \hat{S}_1$ is a cause of some element $e^2 \in S_1$. If $e^2 = e^1$ we are done. If not, we can continue the chain from e^2 until we return to an element $e^3 \in S_1$. If $e^3 = e^1$ or $e^3 = e^2$ we are done; otherwise we continue. Since S_1 is finite, we must eventually return to an element of S_1 we already considered, establishing a causal loop.
2. *Proof for ACL6* Applying Corollary IV.3 to the first condition of ACL6 (Definition VI.8) we have that for every RV $e^1 \in s_1$, there exists an RV $e_2 \in S_2$ such that e^1 is a cause of e_2. Applying the Corollary IV.3 to the second condition, we have that $e_2 \in S_2 \subseteq \hat{S}_2$ must be a cause of some element $e^2 \in s_1$. Either $e^1 = e^2$ and we are done or we continue the chain as in the proof for ACL5.

□

Lemma VI.1. *Let S be a set of RVs and $\mathscr{A}$ be a set of affects relations over them.*

1. *The absence of affects causal loops (Definition VI.1) in $\mathscr{A}$ is a sufficient condition for the existence of a non-trivial embedding of S in a space-time that $\mathscr{A}$ is compatible with.*
2. *If $\mathscr{A}$ is assumed to be a set of affects relations associated with a faithful causal model, then all causal loops are Type 1 affects causal loops and the existence of a non-trivial space-time embedding of S that $\mathscr{A}$ is compatible with is both necessary and sufficient to rule out all causal loops and guarantee the acyclicity of the causal model that generates $\mathscr{A}$.*

Proof. 1. By definition, any set of affects relations that does not contain an affects causal loop is such that the cyclicity of the underlying causal structure is not guaranteed by the affects relations. In other words, it is possible to have the same set of affects relations in a causal model with an acyclic causal structure $\mathcal{G}$. Every causal model over an acyclic causal structure admits a non-trivial space-time embedding since an acyclic causal structure is a directed acyclic graph (DAG) and every DAG implies a partial order. This embedding would be such that the causal arrows $\rightsquigarrow$ of $\mathcal{G}$ flow from past to future in the embedded space-time, which ensures no signalling outside the space-time's future.

2. In faithful causal models, any RV X is a cause of an RV Y if and only if X affects Y. [This follows because if X is a cause of Y then $X \not\perp\!\!\!\perp Y$ in $\mathcal{G}_{\text{do}(X)}$. Since faithful, $X \not\perp\!\!\!\perp Y$ in $\mathcal{G}_{\text{do}(X)}$ and then Lemma IV.2 gives X affects Y. The converse is Lemma IV.3 (which does not rely on faithfulness).] The existence of a causal loop between X and Y corresponds to X being a cause of Y and Y being a cause of X which is equivalent to X affects Y and Y affects X. The latter is the definition of a Type 1 affects causal loop (Definition VI.3). Hence, under the faithfulness assumption, the absence of a Type 1 ACL is equivalent to the acyclicity of the underlying causal structure. As argued in part 1 above, any acyclic causal structure can be non-trivially and compatibly embedded in any space-time structure.

□

4. Proofs of Lemmas A.1 and A.2

Lemma A.1. *For any pairwise disjoint subsets X, Y, Z and W of the observed nodes S of a causal model, we have*

1. $(XZW \perp^d Y)_{\mathcal{G}_{\text{do}(XZ)}} \Rightarrow XZ$ *does not affect* Y *given* W.
2. $(XZW \perp^d Y)_{\mathcal{G}_{\text{do}(XZ)}} \Rightarrow X$ *does not affect* Y *given* $\{\text{do}(Z), W\}$.

Proof. 1. We use Definition IV.1 on the d-separation relation $(XZW \perp^d Y)_{\mathcal{G}_{\text{do}(XZ)}}$ to obtain the conditional independence

$$P_{\mathcal{G}_{\text{do}(XZ)}}(Y|X,Z,W) = P_{\mathcal{G}_{\text{do}(XZ)}}(Y). \tag{D13}$$

Then noting that $(XZW \perp^d Y)_{\mathcal{G}_{\text{do}(XZ)}}$ implies $(XZ \perp^d Y)_{\mathcal{G}_{\text{do}(XZ)}}$, we can apply Corollary IV.2 to the latter d-separation relation to obtain $P_{\mathcal{G}_{\text{do}(XZ)}}(Y) = P_{\mathcal{G}}(Y)$. Combined with the above equation, this gives

$$P_{\mathcal{G}_{\text{do}(XZ)}}(Y|X,Z,W) = P_{\mathcal{G}}(Y). \tag{D14}$$

Now, we show that $P_{\mathcal{G}}(Y) = P_{\mathcal{G}}(Y|W)$ must hold in this case, which would (using the above equation) imply that XZ does not affect Y given W. Suppose by contradiction that $P_{\mathcal{G}}(Y) \neq P_{\mathcal{G}}(Y|W)$, which would imply that $(Y \not\perp^d W)_{\mathcal{G}}$. The assumed d-separation $(XZW \perp^d Y)_{\mathcal{G}_{\text{do}(XZ)}}$ implies that $(W \perp^d Y)_{\mathcal{G}_{\text{do}(XZ)}}$. The only way that we could have d-connection between Y and W in $\mathcal{G}$ but not in $\mathcal{G}_{\text{do}(XZ)}$ is through the existence of a directed path between XZ and Y in $\mathcal{G}$ which gives $(XZ \perp^d Y)_{\mathcal{G}_{\text{do}(XZ)}}$, contradicting our assumption $(XZW \perp^d Y)_{\mathcal{G}_{\text{do}(XZ)}}$. This establishes the first part.

2. We show that $(XZW \perp^d Y)_{\mathcal{G}_{\text{do}(XZ)}}$ implies $(ZW \perp^d Y)_{\mathcal{G}_{\text{do}(Z)}}$, which in turn implies that Z does not affect Y given W. Then along with the first part of the lemma, this gives us $(XZW \perp^d Y)_{\mathcal{G}_{\text{do}(XZ)}} \Rightarrow XZ$ does not affect Y given W and Z does not affect Y given W. Then using Lemma IV.4, this implies that X *does not affect* Y given $\{\text{do}(Z), W\}$, which is the required conclusion.

   Suppose that $(XZW \perp^d Y)_{\mathcal{G}_{\text{do}(XZ)}}$ but $(ZW \not\perp^d Y)_{\mathcal{G}_{\text{do}(Z)}}$. There are two ways that this is possible

   (i) $(Z \not\perp^d Y)_{\mathcal{G}_{\text{do}(Z)}}$: By assumption, we have $(XZW \perp^d Y)_{\mathcal{G}_{\text{do}(XZ)}}$, which implies $(Z \perp^d Y)_{\mathcal{G}_{\text{do}(XZ)}}$. The only way we can then have $(Z \not\perp^d Y)_{\mathcal{G}_{\text{do}(Z)}}$ is through the existence of a directed path from X to Y in $\mathcal{G}_{\text{do}(Z)}$. This gives $(X \perp^d Y)_{\mathcal{G}_{\text{do}(XZ)}}$, which contradicts our assumption.

   (ii) $(W \not\perp^d Y)_{\mathcal{G}_{\text{do}(Z)}}$: The assumption $(XZW \perp^d Y)_{\mathcal{G}_{\text{do}(XZ)}}$ implies $(W \perp^d Y)_{\mathcal{G}_{\text{do}(XZ)}}$. If the d-connection $(W \not\perp^d Y)_{\mathcal{G}_{\text{do}(Z)}}$ is due to a directed path from W to Y in $\mathcal{G}_{\text{do}(Z)}$, this path must go through X in order to ensure that $(W \perp^d Y)_{\mathcal{G}_{\text{do}(XZ)}}$. However, this would violate the original assumption $(XZW \perp^d Y)_{\mathcal{G}_{\text{do}(XZ)}}$ as it would lead to a directed path from X to Y in $\mathcal{G}_{\text{do}(XZ)}$. On the other hand, if the d-connection $(W \not\perp^d Y)_{\mathcal{G}_{\text{do}(Z)}}$ is due to a common cause, it is not possible to have the d-connection $(W \perp^d Y)_{\mathcal{G}_{\text{do}(XZ)}}$, which also contradicts the assumed d-separation.

   The above establishes that $(XZW \perp^d Y)_{\mathcal{G}_{\text{do}(XZ)}}$ implies $(ZW \perp^d Y)_{\mathcal{G}_{\text{do}(Z)}}$, and $(ZW \perp^d Y)_{\mathcal{G}_{\text{do}(Z)}}$ implies Z does not affect Y given W follows from the first part of the proof (with Z playing the role of XZ).

□

Lemma A.2. *For any pairwise disjoint subsets X, Y, Z and W of the observed nodes S of a causal model, we have*

1. $(XZW \not\perp Y)_{\mathcal{G}_{\mathrm{do}(XZ)}} \Rightarrow XZ$ *affects* Y *given* W.

2. $(XZW \not\perp Y)_{\mathcal{G}_{\mathrm{do}(XZ)}}$ and $(ZW \perp^d Y)_{\mathcal{G}_{\mathrm{do}(Z)}} \Rightarrow X$ affects Y *given* $\{\mathrm{do}(Z), W\}$.

Proof. 1. The given dependence $(XZW \not\perp Y)_{\mathcal{G}_{\mathrm{do}(XZ)}}$ is equivalent to

$$\exists x, x', y, z, z', w, w' \text{ s.t. } P_{\mathcal{G}_{\mathrm{do}(XZ)}}(Y = y | X = x, Z = z, W = w) \neq P_{\mathcal{G}_{\mathrm{do}(XZ)}}(Y = y | X = x', Z = z', W = w') \tag{D15}$$

Suppose that XZ does not affect Y given W i.e.,

$$P_{\mathcal{G}_{\mathrm{do}(XZ)}}(Y = y | X = x, Z = z, W = w) = P_{\mathcal{G}}(Y = y | W = w) \qquad \forall x, y, z, w \tag{D16}$$

It is not possible to satisfy both of these equations and $(XZW \not\perp Y)_{\mathcal{G}_{\mathrm{do}(XZ)}}$ must imply XZ affects Y given W.

2. Firstly, the d-separation $(ZW \perp^d Y)_{\mathcal{G}_{\mathrm{do}(Z)}}$ implies that Z does not affect Y given W, which follows from part 1. of Lemma A.1. From part 1. above, we have $(XZW \not\perp Y)_{\mathcal{G}_{\mathrm{do}(XZ)}}$ implies XZ affects Y given W. We now show that Z does not affect Y given W and XZ affects Y given W implies that X affects Y given $\{\mathrm{do}(Z), W\}$, which would complete the proof. Writing out the first two conditions, we have

$$P_{\mathcal{G}_{\mathrm{do}(Z)}}(Y | Z, W) = P_{\mathcal{G}}(Y | W), \tag{D17}$$

$$P_{\mathcal{G}_{\mathrm{do}(XZ)}}(Y | X, Z, W) \neq P_{\mathcal{G}}(Y | W). \tag{D18}$$

Together, these imply that $P_{\mathcal{G}_{\mathrm{do}(XZ)}}(Y | X, Z, W) \neq P_{\mathcal{G}_{\mathrm{do}(Z)}}(Y | Z, W)$ i.e., X affects Y given $\{\mathrm{do}(Z), W\}$.

□

5. Proof of Theorem B.1

Theorem B.1. *Any set of affects relations $\mathscr{A}$ containing an affects causal loops of Type 7 or Type 8 can only arise from a causal model over a cyclic causal structure.*

Proof. The proofs for ACL7 and ACL8 are similar. We describe the proof for ACL8 here, and at the end explain how it also applies to ACL7. Applying Corollary IV.3 to the affects relations $\{S_2'$ affects S_3, S_3' affects S_4, ... ,S_{n-1}' affects S_n, S_n' affects $s_1\} \subseteq \mathscr{A}$ in the second condition of ACL8 (Definition B.2) we have that for each element $e_2' \in S_2$ there exists an element $e_3 \in S_3$ of which it is a cause, for each element $e_3' \in S_3$ there exist an element $e_4 \in S_4$ of which it is a cause, ... , for each element $e_n' \in S_n$ there exist an element $e_1 \in s_1 \subseteq S_1$ of which it is a cause. This does not immediately imply that there is a directed path from S_2' to s_1, since for example the element $e_3' \in S_3$ of which $e_2' \in S_2$ is a cause might not belong to the next set S_3' in the chain, i.e., we could have $e_3 \in S_3 \backslash (S_3 \bigcap S_3')$ if (S_3, S_3') forms an incomplete node of $\mathscr{C}_{e_2}$. In this case, the third condition of Definition B.2 tells us that there is another complete affects chain $\mathscr{D}_{e_2}^{\mathscr{C}}$ that connects $S_3 \backslash (S_3 \bigcap S_3')$ to S_3. Since this is a complete affects chain, we can apply the same argument as in the proof of Theorem VI.1 to conclude that for each element in $S_3 \backslash (S_3 \bigcap S_3')$, there exists an element $e_3^* \in S_3$ of which it is a cause. We consider two cases depending on whether we have $e_3^* \in S_3 \backslash (S_3 \bigcap S_3')$ or $e_3^* \in S_3 \bigcap S_3'$. We will show that in the former case, the affects relations in the secondary chain $\mathscr{D}_{e_2}^{\mathscr{C}}$ already guarantees cyclicity while the latter case, these (the set formed by such secondary chains, one for every incomplete node) guarantee cyclicity when taken together with those in the primary chain $\mathscr{C}_{e_2}$.

In the first case, $\mathscr{D}_{e_2}^{\mathscr{C}}$ corresponds to a Type 5 affects causal loop since it involves a complete chain of irreducible affects relations from a set $S_3 \backslash (S_3 \bigcap S_3')$ on to itself. The cyclicity claim for this case then follows from Theorem VI.1. Therefore, we now consider the case where for each incomplete node (S_i, S_i') of $\mathscr{C}_{e_2}$, the corresponding element $e_i^* \in S_i$ belongs to the intersection of the sets $S_i \bigcap S_i'$. Then, applying Corollary IV.3 repeatedly to each pair of affects relations in $\{S_2'$ affects S_3, S_3' affects S_4, ... ,S_{n-1}' affects S_n, S_n' affects $s_1\} \subseteq \mathscr{A}$, we can conclude that for every element $e_2' \in S_2'$, there exists an element $e_1 \in s_1 \subseteq S_1$ such that e_2' is a cause of e_1. By Definition B.2 (second condition), we considered such a set $\mathscr{A}$ of affects relations for every element $e_2 \in S_2$, defining S_2' such that it contains e_2. Since the above argument holds for all sets of affects relations $\mathscr{A}$ defined as above and for all elements of S_2', this implies that for every element $e_2 \in S_2$, there exists a corresponding element $e_1 \in S_1$ of which it is a cause. Applying Corollary IV.3 to the first condition of Definition B.2 i.e., the irreducible affects relation S_1 affects S_2, we have that for every element $e_1 \in S_1$, there exists a corresponding element $e_2 \in S_2$ of which it is a cause. This was also the case for ACL1-6 as shown in Theorem VI.1, so the statement of the present theorem then follows from the proof of Theorem VI.1.

For ACL7, the first condition says that there is an irreducible affects relation S_1 affects S_2 in $\mathscr{A}$ and the second condition of Definition B.1 guarantees the existence of an affects chain from $s_2 \subseteq S_2$ to S_1. The subtlety here is to note that if $s_2 \subset S_2$, then (S_2, S_2') will be an incomplete node of $\mathscr{C}_{s_2}$. By the above proof for ACL8, we have concluded that the affects relations $\{S_2'$ affects S_3, S_3' affects S_4, $\ldots$,S_{n-1}' affects S_n, S_n' affects $s_1\} \subseteq \mathscr{A}$ along with the third condition of ACL8 (which is similar for ACL7) either imply cyclicity of the causal structure or that for every element $e_2' \in S_2'$, there exists an element $e_1 \in s_1 \subseteq S_1$ such that e_2' is a cause of e_1. If the node (S_2, S_2') is also incomplete as noted above, one can extend the same arguments using the third condition to conclude that either the causal structure is cyclic or for every element $e_2 \in S_2$, there exists an element $e_1 \in s_1 \subseteq S_1$ such that e_2 is a cause of e_1. The same condition was obtained at the end of the previous paragraph, in the proof for ACL8, and shown to imply cyclicity. Therefore this establishes the theorem also for ACL7. □

[1] Vilasini, V. & Colbeck, R. Impossibility of superluminal signalling in Minkowski space-time does not rule out causal loops. arXiv:2206.12887 (2021). link.
[2] Pearl, J. Causality: Models, reasoning, and inference. *Second edition, Cambridge University Press* (2009). link.
[3] Spirtes, P., Glymour, C. N. & Scheines, R. *Causation, prediction, and search* (The MIT Press, 2nd ed., 2001). link.
[4] Kleinberg, S. & Hripcsak, G. A review of causal inference for biomedical informatics. *Journal of Biomedical Informatics* **44**, 1102–1112 (2011). link.
[5] Raita, Y., Camargo, C. A., Liang, L. & Hasegawa, K. Big data, data science, and causal inference: A primer for clinicians. *Frontiers in Medicine* **8**, 998 (2021). link.
[6] Spirtes, P. Graphical models, causal inference, and econometric models. *J. Econ. Methodol.* **12**, 3–34 (2005). link.
[7] Petersen, M. J., Maya L.; van der Laan. Causal models and learning from data. *Epidemiology* **25**, 418–426 (2014). link.
[8] Arti, S., Hidayah, I. & Kusumawardani, S. S. Research trend of causal machine learning method: A literature review. *IJID (International Journal on Informatics for Development)* **9**, 111–118 (2020). link.
[9] Liu, T., Ungar, L. & Kording, K. Quantifying causality in data science with quasi-experiments. *Nature Computational Science* **1**, 24–32 (2021). link.
[10] Bell, J. S. *Speakable and unspeakable in quantum mechanics* (Cambridge University Press, 1987).
[11] Wood, C. J. & Spekkens, R. W. The lesson of causal discovery algorithms for quantum correlations: causal explanations of Bell-inequality violations require fine-tuning. *New Journal of Physics* **17**, 33002 (2015). link.
[12] Tucci, R. R. Quantum Bayesian nets. *International Journal of Modern Physics B* **09**, 295–337 (1995). link.
[13] Leifer, M. S. Quantum dynamics as an analog of conditional probability. *Physical Review A* **74** (2006). link.
[14] Laskey, K. B. Quantum causal networks. arXiv:0710.1200 (2007). link.
[15] Leifer, M. & Poulin, D. Quantum graphical models and belief propagation. *Annals of Physics* **323**, 1899–1946 (2008). link.
[16] Leifer, M. S. & Spekkens, R. W. Towards a formulation of quantum theory as a causally neutral theory of Bayesian inference. *Physical Review A* **88**, 052130 (2013). link.
[17] Henson, J., Lal, R. & Pusey, M. Theory-independent limits on correlations from generalized Bayesian networks. *New Journal of Physics* **16**, 113043 (2014). link.
[18] Pienaar, J. & Brukner, Č. A graph-separation theorem for quantum causal models. *New Journal of Physics* **17**, 73020 (2015). link.
[19] Ried, K. *et al.* A quantum advantage for inferring causal structure. *Nature Physics* **11**, 414–420 (2015). link.
[20] Costa, F. & Shrapnel, S. Quantum causal modelling. *New Journal of Physics* **18**, 63032 (2016). link.
[21] Fritz, T. Beyond Bell's theorem II: Scenarios with arbitrary causal structure. *Communications in Mathematical Physics* **341**, 391–434 (2015). link.
[22] Allen, J.-M. A., Barrett, J., Horsman, D. C., Lee, C. M. & Spekkens, R. W. Quantum common causes and quantum causal models. *Physical Review X* **7**, 031021 (2017). link.
[23] Barrett, J., Lorenz, R. & Oreshkov, O. Quantum causal models. arXiv:1906.10726 (2020). link.
[24] Pienaar, J. Quantum causal models via quantum Bayesianism. *Physical Review A* **101** (2020). link.
[25] Forré, P. & Mooij, J. M. Markov properties for graphical models with cycles and latent variables. arXiv:1710.08775 (2017). link.
[26] Bongers, S., Forré, P., Peters, J. & Mooij, J. M. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics* **49** (2021). link.
[27] Araújo, M., Guérin, P. A. & Baumeler, A. Quantum computation with indefinite causal structures. *Physical Review A* **96** (2017). link.
[28] Barrett, J., Lorenz, R. & Oreshkov, O. Cyclic quantum causal models. arXiv:2002.12157 (2020). link.
[29] Vilasini, V. & Renner, R. Embedding cyclic causal structures in acyclic spacetimes: no-go results for process matrices. arXiv:2203.11245 (2022). link.
[30] Grunhaus, J., Popescu, S. & Rohrlich, D. Jamming nonlocal quantum correlations. *Phys. Rev. A* **53**, 3781–3784 (1996). link.
[31] Horodecki, P. & Ramanathan, R. The relativistic causality versus no-signaling paradigm for multi-party correlations. *Nature Communications* **10**, 1701 (2019). link.
[32] Oreshkov, O., Costa, F. & Brukner, Č. Quantum correlations with no causal order. *Nature Communications* **3**, 1092 (2012). link.
[33] Zych, M., Costa, F., Pikovski, I. & Brukner, Č. Bell's theorem for temporal order. *Nature Communications* **10**, 3772 (2019). link.
[34] Agresti, I. *et al.* Experimental test of quantum causal influences. arXiv:2108.08926 (2021). link.
[35] Kent, A. Secure classical bit commitment over finite channels. *Journal of Cryptology* **18**, 313–335 (2005). link.
[36] Colbeck, R. & Kent, A. Variable bias coin tossing. *Physical Review A* **73**, 032320 (2006).
[37] Bohm, D. A suggested interpretation of the quantum theory in terms of "hidden" variables. I. *Physical Review* **85**, 166–179 (1952). link.
[38] Vilasini, V. & Colbeck, R. *In preparation* (2022).
[39] Barrett, J. Information processing in generalized probabilistic theories. *Phys. Rev. A* **75**, 032304 (2007). link.
[40] Brunner, N., Cavalcanti, D., Pironio, S., Scarani, V. & Wehner, S. Bell nonlocality. *Reviews of Modern Physics* **86**,

419–478 (2014). link.
[41] Geiger, D. Towards the formalization of informational dependencies. *Tech. rep. 880053. UCLA Computer Science* (1987). link.
[42] Verma, T. & Pearl, J. Causal networks: Semantics and expressiveness. In *Proceedings of the Fourth Annual Conference on Uncertainty in Artificial Intelligence*, UAI '88, 69–78 (North-Holland Publishing Co., Amsterdam, The Netherlands, The Netherlands, 1990). link.
[43] Pearl, J. & Dechter, R. Identifying independencies in causal graphs with feedback. arXiv:1302.3595 (2013). link.
[44] Friedman, M. The fed's thermostat. *The Wall Street Journal* (2003). link.
[45] Rowe, N. Why there's so little good evidence that fiscal (or monetary) policy works (online) (2009). link.
[46] Brulé, J. A causation coefficient and taxonomy of correlation/causation relationships. arXiv:1708.05069 (2017). link.
[47] Colbeck, R. & Renner, R. No extension of quantum theory can have improved predictive power. *Nature Communications* **2**, 411 (2011). link.
[48] Colbeck, R. & Renner, R. A short note on the concept of free choice. arXiv:1302.4446 (2013). link.
[49] Salazar, R. *et al.* A no-go theorem for device-independent security in relativistic causal theories. arXiv:1712.01030 (2020). link.
[50] Reichenbach, H. *The direction of time* (Univ. of California Press, Berkeley - Los Angeles, 1956).
[51] Pearl, J. Causal diagrams for empirical research. *Biometrika* **82**, 669–688 (1995). link.
[52] Kempf, A. Replacing the notion of spacetime distance by the notion of correlation. *Frontiers in Physics* **9** (2021). link.
[53] Bombelli, L., Lee, J., Meyer, D. & Sorkin, R. D. Space-time as a causal set. *Phys. Rev. Lett.* **59**, 521–524 (1987). link.
[54] Surya, S. The causal set approach to quantum gravity. *Living Reviews in Relativity* **22**, 5 (2019). link.
[55] Dukovski, I. Causal structure of spacetime and geometric algebra for quantum gravity. *Phys. Rev. D* **87**, 064022 (2013). link.
[56] Coecke, B. & Kissinger, A. *Picturing Quantum Processes: A First Course in Quantum Theory and Diagrammatic Reasoning* (Cambridge University Press, 2017). link.
[57] Giacomini, F., Castro-Ruiz, E. & Brukner, Č. Quantum mechanics and the covariance of physical laws in quantum reference frames. *Nature Communications* **10**, 494 (2019). link.
[58] Castro-Ruiz, E., Giacomini, F., Belenchia, A. & Brukner, Č. Quantum clocks and the temporal localisability of events in the presence of gravitating quantum systems. *Nature Communications* **11** (2020). link.
[59] Deutsch, D. Quantum mechanics near closed timelike lines. *Phys. Rev. D* **44**, 3197–3217 (1991). link.
[60] Bennett, C. & Schumacher, B. Talk at QUPON Wien. link.
[61] Svetlichny, G. Time travel: Deutsch vs. teleportation. *International Journal of Theoretical Physics* **50**, 3903–3914 (2011). link.
[62] Lloyd, S., Maccone, L., Garcia-Patron, R., Giovannetti, V. & Shikano, Y. Quantum mechanics of time travel through post-selected teleportation. *Phys. Rev. D* **84**, 025007 (2011). link.
[63] Lloyd, S. *et al.* Closed timelike curves via postselection: Theory and experimental test of consistency. *Phys. Rev. Lett.* **106**, 040403 (2011). link.
[64] Aaronson, S. & Watrous, J. Closed timelike curves make quantum and classical computing equivalent. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **465**, 631–647 (2008). link.
[65] Aharonov, Y., Bergmann, P. G. & Lebowitz, J. L. Time symmetry in the quantum process of measurement. *Phys. Rev.* **134**, B1410–B1416 (1964). link.
[66] Oeckl, R. A local and operational framework for the foundations of physics. *Advances in Theoretical and Mathematical Physics* **23**, 437–592 (2019). link.
[67] Di Biagio, A., Donà, P. & Rovelli, C. The arrow of time in operational formulations of quantum theory. *Quantum* **5**, 520 (2021). link.
[68] Hardy, L. Time symmetry in operational theories. arXiv:2104.00071 (2021). link.
[69] Chiribella, G. & Liu, Z. Quantum operations with indefinite time direction. arXiv:2012.03859 (2021). link.
[70] Aharonov, Y. & Vaidman, L. Complete description of a quantum system at a given time. *Journal of Physics A: Mathematical and General* **24**, 2315–2328 (1991). link.
[71] Silva, R. *et al.* Pre- and postselected quantum states: Density matrices, tomography, and Kraus operators. *Phys. Rev. A* **89**, 012121 (2014). link.
[72] Zhalama, Zhang, J. & Mayer, W. Weakening faithfulness: some heuristic causal discovery algorithms. *International Journal of Data Science and Analytics* **3**, 93–104 (2017). link.
[73] Chiribella, G., D'Ariano, G. M., Perinotti, P. & Valiron, B. Quantum computations without definite causal structure. *Physical Review A* **88**, 022318 (2013). link.
[74] Procopio, L. M. *et al.* Experimental superposition of orders of quantum gates. *Nature Communications* **6**, 7913 (2015). link.
[75] Rubino, G. *et al.* Experimental verification of an indefinite causal order. *Science Advances* **3** (2017). link.
[76] Portmann, C., Matt, C., Maurer, U., Renner, R. & Tackmann, B. Causal boxes: Quantum information-processing systems closed under composition. *IEEE Transactions on Information Theory* **63**, 3277–3305 (2017). link.
[77] Vilasini, V. Approaches to causality and multi-agent paradoxes in non-classical theories. *PhD Thesis, University of York* (2021). link.
[78] Neal, R. M. On deducing conditional independence from d-separation in causal graphs with feedback (research note). *Journal of Artificial Intelligence Research* **12**, 87–91 (2000). link.